A Maximum Likelihood Approach to Masking-based Speech Enhancement Using Deep Neural Network

Qing Wang[1], *Jun Du*[1], *Li Chai*[1], *Li-Rong Dai*[1], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Georgia Institute of Technology, Atlanta, GA. USA
{xiaosong, cl122}@mail.ustc.edu.cn, {jundu, lrdai}@ustc.edu.cn, chl@ece.gatech.edu

Abstract

The minimum mean squared error (MMSE) is usually adopted as the training criterion for speech enhancement based on deep neural network (DNN). In this study, we propose a probabilistic learning framework to optimize the DNN parameter for masking-based speech enhancement. Ideal ratio mask (IRM) is used as the learning target and its prediction error vector at the DNN output is modeled to follow statistically independent generalized Gaussian distribution (GGD). Accordingly, we present a maximum likelihood (ML) approach to DNN parameter optimization. We analyze and discuss the effect of shape parameter of GGD on noise reduction and speech preservation. Experimental results on the TIMIT corpus show the proposed ML-based learning approach can achieve consistent improvements over MMSE-based DNN learning on all evaluation metrics. Less speech distortion is observed in ML-based approach especially for high frequency units than MMSE-based approach.

Index Terms: speech enhancement, deep neural network, ideal ratio mask, the prediction error, generalized Gaussian distribution, maximum likelihood estimation

1. Introduction

Single-channel speech enhancement aims to improve the quality and intelligibility of a speech signal degraded by adverse noise surroundings and plays an important role in real-world applications, such as automatic speech recognition (ASR), mobile speech communication, and hearing aids [1]. In the past several decades, unsupervised speech enhancement techniques, such as spectral subtraction [2], Wiener filtering [3], minimum mean squared error (MMSE) estimation [4] and optimally-modified log-spectral amplitude (OM-LSA) speech estimator [5] have been studied extensively. However, most of these techniques fail to track non-stationary noise environments and often cause musical noise artifacts.

Recently, with the fast development of deep learning techniques [6, 7], supervised machine learning has attracted much attention. According to the definition of the learning target, supervised speech enhancement methods can be categorized into (i) mapping-based methods and (ii) masking-based methods [8]. Mapping-based methods directly map clean speech from a noisy signal. For example, Xu [9] proposed deep neural network (DNN) based speech enhancement framework to map noisy log-power spectra (LPS) features [10] to clean LPS features. In [11], deep denoising autoencoder was adopted to model the complicated relationship between noisy speech and clean speech. More complex neural network architectures, such as recurrent neural network (RNN) [12] and long short-term memory (LSTM) RNN [13], were designed for speech enhancement to achieve performance improvements.

Masking-based methods firstly learn a time-frequency (T-F) mask from a noisy signal, and then use the estimated mask to predict the clean speech. A DNN was used to predict ideal binary mask (IBM) for speech separation [14]. Wang [15] used different training targets for speech separation and suggested that ideal ratio mask (IRM) outperformed IBM in terms of objective intelligibility and quality metrics. A single DNN to jointly predict the real and imaginary components of the complex ideal ratio mask (cIRM), was adopted in [16] and it was suggested that cIRM should be preferred over the conventional magnitude-only IRM. Huang [17] adopted deep recurrent neural network (DRNN) to jointly optimize the T-F masking functions with the deep learning model.

In DNN-based speech enhancement methods, the optimization of objective function is based on MMSE criterion. However, the MMSE-based estimation method is not very robust in adverse acoustic scenarios, which may cause additional speech distortion due to the over-smoothing problem. This effects and limits the quality and intelligibility of the denoised speech. Researchers began to explore new objective functions. Shivakumar [18] proposed a novel objective loss function, which took into account the perceptual quality of speech. A weighted reconstruction loss function was introduced into the traditional denoising autoencoder model in [19]. Kinoshita [20] proposed a mixture density network to map a set of Gaussian mixture model (GMM) parameters representing the distribution of a target variable from an input feature. Koizumi [21] proposed a training method for DNN-based source enhancement to increase objective sound quality assessment (OSQA) scores. Erdogan [22] developed a phase-sensitive objective function for speech separation. In our recent work [23, 24, 25], a maximum likelihood (ML) criterion was used to train mapping-based DNNs for speech enhancement, speech separation, and speech dereverberation. Compared with MMSE-based approach, the ML-based approach could achieve better convergence and objective metrics.

As a measure to estimate the speech presence in a local T-F unit, IRM has been widely used in speech recognition [26, 27], speech enhancement [28], and speech separation [29]. Wang [15] suggested to use mask as training target because its normalization form could reduce the dynamic range of target values and thus got different training efficiency compared to mapping. The effectiveness of IRM in speech enhancement has been demonstrated in our recent work [30, 31]. It is known that generalized Gaussian distribution (GGD) [32] is used to model the probability density function (PDF) of a signal. In this study, we explore the ML solution within the probabilistic learning framework to optimize masking-based DNN parameter. Under the assumption that each dimension of the IRM prediction error vector at the DNN output follows GGD, a training procedure is

978-1-5386-5627-3/18/$31.00 ©2018 IEEE

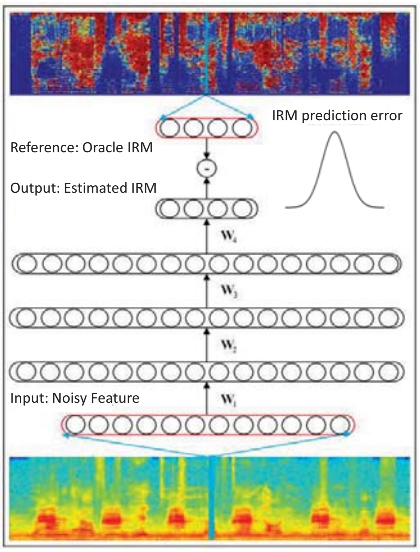


Figure 1: *The ML-DNN architecture for speech enhancement.*

designed to update the DNN parameter. We analyze and discuss the effect of shape parameter of GGD on evaluation metrics for ML-based DNN (ML-DNN) approach. The experiments evaluated on unseen noise types show that the proposed ML-DNN approach outperforms MMSE-based DNN (MMSE-DNN) approach for all objective evaluation metrics.

2. The Proposed ML-DNN Approach

In this study, we redefine the objective function in the probabilistic learning framework and adopt the maximum likelihood estimation to update the parameter of masking-based DNN, as shown in Figure 1. The input of DNN is the $(2\tau + 1)D$-dimensional LPS feature vector of noisy speech with an acoustic context of $2\tau + 1$ neighbouring frames while the output is the D-dimensional IRM vector. And the reference is the corresponding D-dimensional oracle IRM vector.

In conventional MMSE-DNN, a mini-batch based stochastic gradient descent (SGD) algorithm is adopted to optimize the model parameter using the following loss function,

$$E = \frac{1}{N}\sum_{n=1}^{N} ||\hat{\mathbf{m}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}) - \mathbf{m}_n||_2^2 \tag{1}$$

where E is the mean squared error. $\hat{\mathbf{m}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W})$ and $\mathbf{m}_n$ are the estimated and reference IRM vector at the n^{th} frame, respectively, with N representing the mini-batch size, $\mathbf{y}_{n-\tau}^{n+\tau}$ being the noisy LPS feature vector where the window size of context is $2\tau + 1$, and $\mathbf{W}$ denoting the DNN parameter to be learned. The IRM prediction error vector $\mathbf{e}_n$ at the n^{th} frame could be defined as:

$$\mathbf{e}_n = \mathbf{m}_n - \hat{\mathbf{m}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}) \tag{2}$$

We assume that each dimension of the IRM prediction error vector follows a univariate GGD with a zero mean, an unrestricted scale parameter α_d and a known shape parameter β_d:

$$p(e_{n,d}|\alpha_d, \beta_d) = \frac{\beta_d}{2\alpha_d\Gamma(\frac{1}{\beta_d})}\exp\left(-(\frac{|e_{n,d}|}{\alpha_d})^{\beta_d}\right) \tag{3}$$

Correspondingly, assuming that each dimension of the IRM prediction error vector is drawn independently from the GGDs, we can get a multivariate GGD as follows:

$$p(\mathbf{e}_n|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{D} \frac{\beta_d}{2\alpha_d\Gamma(\frac{1}{\beta_d})}\exp\left(-(\frac{|e_{n,d}|}{\alpha_d})^{\beta_d}\right) \tag{4}$$

where $\boldsymbol{\alpha} = \alpha_d, \boldsymbol{\beta} = \beta_d, d = 1, 2, ..., D$. If the reference IRM vector $\mathbf{m}_n$ is also a random vector, then Eq. (4) is equivalent to:

$$\begin{aligned} &p(\mathbf{m}_n|\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \\ &\prod_{d=1}^{D} \frac{\beta_d}{2\alpha_d\Gamma(\frac{1}{\beta_d})}\exp\left(-(\frac{|m_{n,d} - \hat{m}_{n,d}|}{\alpha_d})^{\beta_d}\right) \end{aligned} \tag{5}$$

Given a mini-batch training set with N data pairs $(\mathbf{Y}, \mathbf{M}) = \{(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{m}_n)|n = 1, 2, ..., N\}$ and making the assumption that these data pairs are drawn independently from the distribution in Eq. (5), we can define the likelihood function as:

$$\begin{aligned} &p(\mathbf{M}|\mathbf{Y}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \\ &\prod_{n=1}^{N}\prod_{d=1}^{D} \frac{\beta_d}{2\alpha_d\Gamma(\frac{1}{\beta_d})}\exp\left(-(\frac{|m_{n,d} - \hat{m}_{n,d}|}{\alpha_d})^{\beta_d}\right) \end{aligned} \tag{6}$$

Accordingly, the log-likelihood function can be written as:

$$\begin{aligned} &\ln p(\mathbf{M}|\mathbf{Y}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \\ &\sum_{n=1}^{N}\sum_{d=1}^{D} \ln\left(\frac{\beta_d}{2\alpha_d\Gamma(\frac{1}{\beta_d})}\exp\left(-(\frac{|m_{n,d} - \hat{m}_{n,d}|}{\alpha_d})^{\beta_d}\right)\right) \end{aligned} \tag{7}$$

If we assume that the distribution of each dimension has the same known shape factor β, we can get the log-likelihood function as follows:

$$\begin{aligned} \ln p(\mathbf{M}|\mathbf{Y}, \mathbf{W}, \boldsymbol{\alpha}, \beta) = &\sum_{n=1}^{N}\sum_{d=1}^{D} \ln\left(\frac{\beta}{2\alpha_d\Gamma(\frac{1}{\beta})}\right) \\ &- \sum_{n=1}^{N}\sum_{d=1}^{D}\left(\frac{|m_{n,d} - \hat{m}_{n,d}|}{\alpha_d}\right)^{\beta} \end{aligned} \tag{8}$$

where the parameter set $(\mathbf{W}, \boldsymbol{\alpha})$ is to be optimized. We adopt maximum likelihood criterion to alternately update $\mathbf{W}$ and $\boldsymbol{\alpha}$. Firstly keep $\mathbf{W}$ fixed maximize Eq.(8) with respect to $\boldsymbol{\alpha}$. The scale parameter is obtained as:

$$\alpha_d = \left(\frac{\beta\sum_{n=1}^{N} |m_{n,d} - \hat{m}_{n,d}|^{\beta}}{N}\right)^{\frac{1}{\beta}} \tag{9}$$

Then to maximize Eq.(8) with respect $\mathbf{W}$ is equivalent to minimize the following loss function:

$$E(\mathbf{W}) = \sum_{n=1}^{N}\sum_{d=1}^{D}\left(\frac{|m_{n,d} - \hat{m}_{n,d}|}{\alpha_d}\right)^{\beta} \tag{10}$$

where $\mathbf{W}$ is optimized by the back-propagation (BP) algorithm with a SGD method in the mini-batch mode with N sample frames. The proposed training procedure is repeated until convergence criterion is satisfied or a maximum number of iterations is exceeded. The whole training procedure is summarized as Algorithm 1.

Algorithm 1 Procedure of ML-DNN training

```
Step 1: Initialization
    Initialize the DNN parameter $\mathbf{W}$ randomly.
Step 2: Alternative optimization in mini-batch mode
    Step 2.1: Fix $\mathbf{W}$ and update $\boldsymbol{\alpha}$ via Eq. (9)
    Step 2.2: Fix $\boldsymbol{\alpha}$ and update $\mathbf{W}$ via Eq. (10)
Step 3: Go to Step 2 for the next mini-batch
```

3. Experimental Results and Analysis

In this study, experiments were conducted on speech waveforms with 16 kHz. 115 noise types including 100 noise types in [33] and some other musical noises were adopted to improve the generalization capacity of DNN. All 4620 utterances from the training set of the TIMIT corpus were corrupted with the above-mentioned 115 noise types at six levels of SNR, i.e., 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB, to build a 80-hour training set. The 192 utterances from core test set of TIMIT corpus were used to construct the test set. 8 unseen noise types (Destroyer Engine, Factory, Military Vehicle, Machine Gun, Pink, Volvo, Speech Babble, and White) from the NOISEX-92 corpus [34] were adopted for testing.

The frame length was set to 512 samples (32 msec) with a frame shift of 256 samples. With short-time Fourier analysis, 257-dimensional LPS features [10] were obtained to train DNNs. Mean and variance normalization were applied to the input feature vectors of the DNN. Sigmoid activation function was employed for all layers. All DNN configurations were fixed at 3 hidden layers, 2048 units for each hidden layer and 7-frame input. DNNs were initialized with random weights. The mini-batch size N was set to 128. The learning rate for fine-tuning was initially 0.1 for the first 10 iterations and decreased by 10% after every iteration in the next 40 iterations. The momentum rate was 0.9 and the weight decay coefficient was 0.00001.

We adopted four objective metrics to evaluate the performance of our proposed ML-DNN. A perceptual evaluation of speech quality (PESQ) [35] and the short-time objective intelligibility (STOI, in %) [36] were used to assess the quality and intelligibility of enhanced speech. Segmental SNR (SSNR) measures the degree of noise reduction while log-spectral distortion (LSD) is designed as an indicator of the speech distortion [10].

3.1. Evaluation on objective metrics

Table 1 and 2 show comparisons of the average performance on the test set by MMSE-based approach and ML-based approach (PESQ and STOI) and (SSNR and LSD), respectively, at four SNR levels across 8 unseen noise types. Different shape parameters of GGD were used for ML-based approach. Firstly, for ML-based approach, the performance was greatly affected by the shape parameter of GGD. ML-DNN which used a larger shape parameter could achieve better PESQ and STOI metrics while ML-DNN which used a smaller shape parameter could achieve better SSNR and LSD metrics. We will analyze and discuss the distributions of IRM prediction error from ML-DNNs trained with GGDs using different shape parameters in detail in the next subsection, which could explain the difference in objective metrics. Secondly, from these two Tables we can see that by setting the shape parameter of GGD to 3, all four evaluation metrics (PESQ, STOI, SSNR, and LSD) could achieve relatively better results overall. Compared to MMSE-DNN, consistently large improvements were achieved for the four evaluation metrics at four SNR levels for ML-DNN trained with GGD using shape parameter 3, with average gains of 0.12, 2.9, 1.57dB, and 1.27dB for PESQ, STOI, SSNR, and LSD, respectively.

Table 1: *Comparison of average PESQ and STOI (in %) metrics by MMSE-based approach and ML-based approach on test sets at four SNR levels across 8 unseen noise types, where β in MLGGDβ represents the shape parameter of GGD for ML-based approach.*

	Approaches	10dB	5dB	0dB	-5dB	Ave
PESQ	MMSE	3.18	2.87	2.52	2.17	2.68
	MLGGD1	3.09	2.70	2.25	1.81	2.46
	MLGGD2	3.26	2.93	2.52	2.10	2.70
	MLGGD3	3.31	3.00	2.65	2.26	2.80
	MLGGD4	3.27	2.96	2.61	2.23	2.77
STOI	MMSE	90.8	85.6	78.3	69.6	81.1
	MLGGD1	93.1	88.0	79.8	68.5	82.3
	MLGGD2	93.4	88.7	81.3	71.1	83.6
	MLGGD3	93.3	88.7	81.7	72.2	84.0
	MLGGD4	93.2	88.7	81.7	72.5	84.0

Table 2: *Comparison of average SSNR and LSD metrics by MMSE-based approach and ML-based approach on test sets at four SNR levels across 8 unseen noise types, where β in MLGGDβ represents the shape parameter of GGD for ML-based approach.*

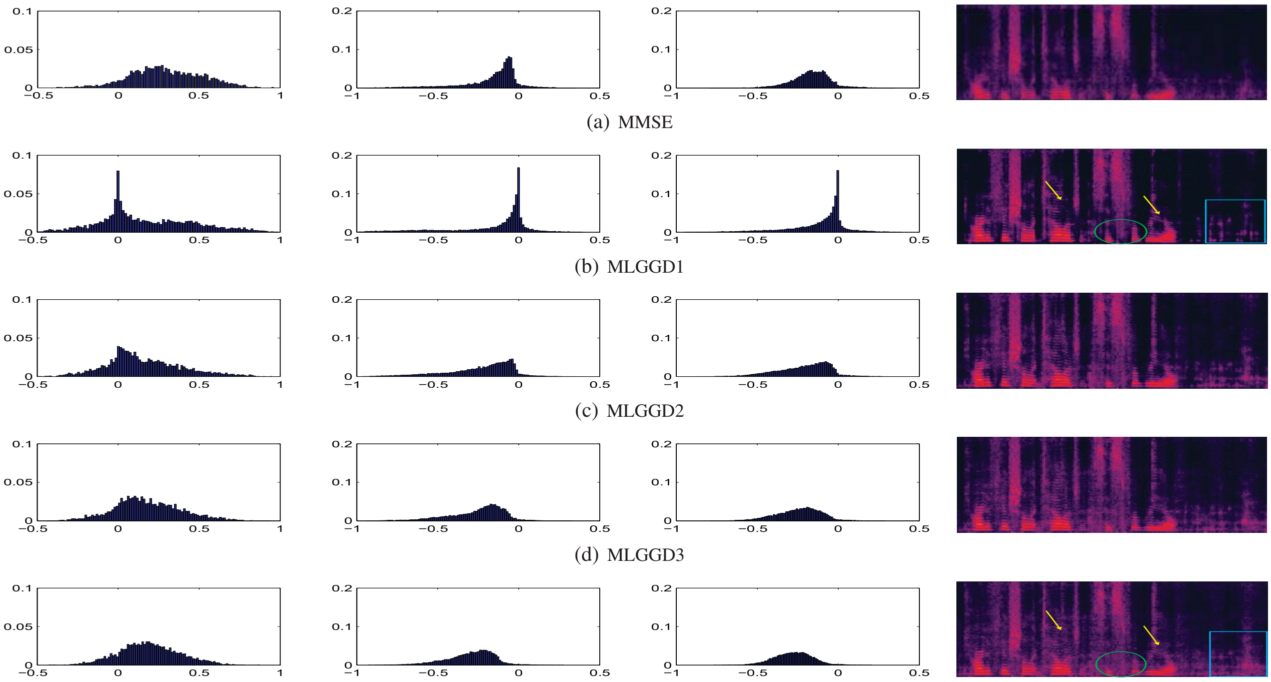
	Methods	10dB	5dB	0dB	-5dB	Ave
SSNR	MMSE	6.87	4.94	3.20	1.71	4.18
	MLGGD1	10.54	8.08	5.75	3.68	7.01
	MLGGD2	10.09	7.63	5.28	3.17	6.54
	MLGGD3	9.14	6.80	4.54	2.52	5.75
	MLGGD4	8.03	5.86	3.74	1.79	4.85
LSD	MMSE	2.91	3.75	4.86	6.26	4.44
	MLGGD1	2.02	2.69	3.48	4.38	3.14
	MLGGD2	2.03	2.66	3.41	4.21	3.10
	MLGGD3	2.17	2.75	3.46	4.32	3.17
	MLGGD4	2.31	2.87	3.57	4.45	3.30

3.2. Statistical analysis on prediction errors

As can be seen from Table 1 and 2, the shape parameters of GGD have a great influence on the final performance, and the influence on speech quality, speech intelligibility, noise reduction, and speech distortion is not consistent. That is, slightly larger shape parameters achieved better speech quality and intelligibility, while slightly smaller shape parameters resulted in better noise reduction and less speech distortion. Therefore, we analyze and discuss the effect of GGD's shape parameter on the distribution of IRM prediction error, which in turn can be related to the trend of the four objective evaluation metrics.

Figure 2 shows the distributions of IRM prediction error and spectrograms from well-trained DNN using MMSE, MLGGD1, MLGGD2, MLGGD3, and MLGGD4 approach for one utterance corrupted by Factory noise at SNR=5dB, where β in MLGGDβ represents the shape parameter of GGD for ML-based approach. To have a deeper understanding of the GGD's shape parameter on four evaluation metrics, a higher resolution analysis as shown in Figure 2 is given by listing the distributions of IRM prediction error on speech-dominant T-F units in low-frequency band, speech-dominant T-F units in high-frequency band, and noise-dominant T-F units in low-frequency band, respectively.

By observing the first column in Figure 2, the distribution of IRM prediction error gradually shifted to the right when the shape parameter of GGD changed from 1 to 4. According to the definition of IRM prediction error in Eq. (2), the rightward shift of the IRM prediction error indicated that the estimated IRM



(a) MMSE

(b) MLGGD1

(c) MLGGD2

(d) MLGGD3

(e) MLGGD4

Figure 2: *The distributions of IRM prediction error and spectrograms from well-trained DNN using MMSE, MLGGD1, MLGGD2, MLGGD3, and MLGGD4 approach respectively for one utterance corrupted by Factory noise at SNR=5dB (from left to right): speech-dominant T-F units in low-frequency band, speech-dominant T-F units in high-frequency band, noise-dominant T-F units in low-frequency band, and the enhanced speech spectrograms.*

gradually became smaller. This means speech distortion in low-frequency band gradually became larger when the shape parameter of GGD changed from 1 to 4 as shown in the green circles in the last column, which was consistent with LSD metric in Table 2. With the gradual increase of shape parameter, the IRM prediction error from ML-DNN shifted gradually to the left for those speech-dominant T-F units in high-frequency band, which indicated the estimated IRM gradually became larger. Thus speech preservation in high-frequency band was much better for a larger shape parameter as indicated by the yellow arrows, and accordingly speech quality and intelligibility were improved. Similarly, the leftward shift of the IRM prediction error caused more residual noise with the gradual increase of shape parameter as shown in the third column and the blue rectangles, which was consistent with the SSNR metric in Table 2.

Figure 3 shows the comparison of spectrograms from MMSE-DNN and ML-DNN (shape parameter of GGD was set to 3) for one utterance corrupted by Factory noise at SNR=5dB. ML-DNN outperformed MMSE-DNN with better speech preservation in both low-frequency and high-frequency bands, as shown in the blue rectangles and the yellow circles. However, ML-DNN might introduce a litter more residual noise as shown in the green dashed rectangles.

4. Conclusion and Future Work

In this paper, we propose a maximum likelihood approach to optimize the parameter set of masking-based DNN for speech enhancement. Based on the assumption that the IRM prediction error vector at the DNN output follows generalized Gaussian distribution, we adopt maximum likelihood criterion to alternately update the DNN parameter and the scale parameter of GGD. We analyze and discuss the effect of shape parameter on objective evaluation metrics. Compared with the conventional MMSE criterion, the ML approach could achieve consistent improvements on four objective evaluation metrics with less speech distortion.

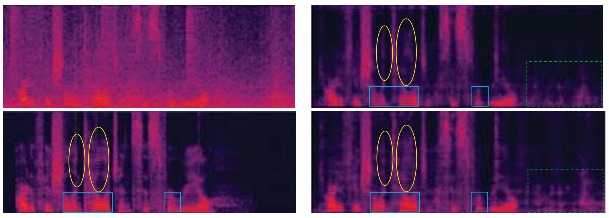


Figure 3: *Comparison of spectrograms for one utterance corrupted by Factory noise at SNR=5dB: noisy speech (upper left), clean speech (bottom left), MMSE-DNN (upper right), ML-DNN where the shape parameter of GGD is 3 (bottom right).*

5. Acknowledgements

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005, MOE-Microsoft Key Laboratory of USTC, and Huawei Noah's Ark Lab.

6. References

[1] J. Benesty, S. Makino, and J. D. Chen, *Speech Enhancement*. Springer, 2005.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[6] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

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

[8] X.-L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 5, pp. 967–977, 2016.

[9] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[10] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Proc. Interspeech*, 2008, pp. 569–572.

[11] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.

[12] A. L. Maas, T. M. ONeil, A. Y. Hannun, and A. Y. Ng, "Recurrent neural network feature enhancement: The 2nd CHiME challenge," in *Proceedings The 2nd CHiME Workshop on Machine Listening in Multisource Environments held in conjunction with ICASSP*, 2013, pp. 79–80.

[13] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[14] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[15] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[16] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *Proc. ICASSP*, 2016, pp. 5220–5224.

[17] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[18] P. G. Shivakumar and P. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement," in *Proc. Interspeech*, 2016, pp. 3743–3747.

[19] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[20] K. Kinoshita, M. Delcroix, A. Ogawa, T. Higuchi, and T. Nakatani, "Deep mixture density network for statistical model-based feature enhancement," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 251–255.

[21] Y. Koizumi, K. Niwa, Y. Hioka, K. Koabayashi, and Y. Haneda, "DNN-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[22] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 708–712.

[23] L. Chai, J. Du, and Y. Wang, "Gaussian density guided deep neural network for single-channel speech enhancement," in *Machine Learning for Signal Processing (MLSP), 2017 IEEE 27th International Workshop on*. IEEE, 2017, pp. 1–6.

[24] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A maximum likelihood approach to deep neural network based nonlinear spectral mapping for single-channel speech separation," *Proc. Interspeech 2017*, pp. 1178–1182, 2017.

[25] X. Wang, J. Du, and Y. Wang, "A maximum likelihood approach to deep neural network based speech dereverberation," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*. IEEE, 2017, pp. 155–158.

[26] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[27] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7092–7096.

[28] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. Wouters, "The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 1, pp. 63–72, 2013.

[29] J. Chen and D. L. Wang, "DNN based mask estimation for supervised speech separation," in *Audio source separation*. Springer, 2018, pp. 207–235.

[30] Q. Wang, J. Du, L. R. Dai, and C. H. Lee, "Joint noise and mask aware training for DNN-based speech enhancement with SUB-band features," in *Hands-free Speech Communication and Microphone Arrays*, 2017, pp. 101–105.

[31] ——, "A multiobjective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 7, pp. 1181–1193, 2018.

[32] M. West, "On scale mixtures of normal distributions," *Biometrika*, vol. 74, no. 3, pp. 646–648, 1987.

[33] G. Hu, "100 nonspeech environmental sounds," http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html, 2014.

[34] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.