INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

KL-divergence Regularized Deep Neural Network Adaptation for Low-resource Speaker-dependent Speech Enhancement

Li Chai[1], *Jun Du*[2], *and Chin-Hui Lee*[3]

[1]School of Data Science, University of Science and technology of China, Hefei, Anhui, P. R. China
[2]University of Science and technology of China, Hefei, Anhui, P. R. China
[3]Georgia Institute of Technology, Atlanta, GA. USA

`cl122@mail.ustc.edu.cn, jundu@ustc.edu.cn, chl@ece.gatech.edu`

Abstract

In this paper, we propose a Kullback-Leibler divergence (KLD) regularized approach to adapting speaker-independent (SI) speech enhancement model based on regression deep neural networks (DNNs) to another speaker-dependent (SD) model using a tiny amount of speaker-specific adaptation data. This algorithm adapts the DNN model conservatively by forcing the conditional target distribution estimated from the SD model to be close to that from the SI model. The constraint is realized by adding KLD regularization to our previously proposed maximum likelihood objective function. Experimental results demonstrate that, even with only 10 seconds of SD adaptation data, the proposed framework consistently achieves speech intelligibility improvements under all 15 unseen noise types evaluated and at all signal-to-noise ratio levels for all 8 test speakers from the WSJ0 evaluation set.

Index Terms: speaker-dependent speech enhancement, deep neural network, maximum likelihood, conditional target distribution, Kullback-Leibler divergence regularization

1. Introduction

Considering the various complicated situation, the speech enhancement performance in real acoustic environments is still unsatisfactory and many problems should be solved. The traditional speech enhancement algorithms, such as spectral subtraction [1–3], Wiener filtering [4, 5], a minimum mean squared error (MMSE) estimator [6, 7] and an optimally-modified log-spectral amplitude speech estimator [8], were developed during the past several decades. However, they often fail to track non-stationary noise for real-world scenarios in unexpected acoustic conditions and are often ineffective in improving speech intelligibility [9, 10].

Recently, deep learning-based speech enhancement has shown considerable success. Xu et al. proposed a regression deep neural network (DNN)-based speech enhancement framework [11, 12] which was adopted to model the complicated relationship between the noisy speech and clean speech features via training a deep and wide neural network architecture using a large collection of heterogeneous training data and the abundant acoustic context information. In addition to direct mapping, masking techniques have been used to enhance speech by making classifications of time-frequency units, such as estimating the ideal binary mask or smoothed ideal ratio mask [13, 14]. In [15], Erdogan et al developed a phase-sensitive mask that incorporates the phase difference between noisy speech and clean speech. To jointly enhance the magnitude and phase spectra, a complex ideal ratio mask was proposed in [16]. However, these DNN-based speech enhancement algorithms still suffer performance degradation under mismatch conditions. In real-world scenarios, the acoustic environment where we deploy our enhancement model can be vastly different from our training examples, and unseen noises and speakers can degrade the quality of processed signal.

For DNN-based speech enhancement, generalization to unseen noises and speakers is a critical issue. Although the generalization capability can be increased by collecting as many types of noises and speakers as possible, it is not practical to cover potentially infinite noise and speaker types that may occur in real acoustic conditions. Personalized services are needed and feasible today. Therefore, it is meaningful to investigate the speaker-dependent (SD) speech enhancement. In [17, 18], a unified DNN-based SD speech separation and enhancement system was proposed to jointly handle both background noise and interfering speech, where the speaker-specific data used for DNN training is about 2 hours. In [19], a two-stage approach was proposed for SD enhancement of far-field microphone array speech collected in reverberant conditions corrupted by interfering speakers and noises, where 5 minutes of speaker-specific data is used. [20] adopted more than 5 minutes of speaker-specific data to train a two-stage single-channel SD speech separation system. However, in many cases, a large amount of speech data is hard to collect for a certain specific speaker in real-world conditions. Accordingly, in this paper, we investigate how to adapt a well-trained speaker-independent (SI) model towards a SD model using a tiny amount of speaker-specific data for DNN-based speech enhancement.

In [21], we proposed a probabilistic learning framework to parameter optimization for DNN-based speech enhancement, where a new objective function is derived according to the maximum likelihood (ML) criterion by characterizing the prediction error vector as a multivariate Gaussian density. Motivated by [22], we propose a Kullback-Leibler divergence (KLD) regularized technique based on the probabilistic learning framework to adapt a well-trained SI model to a SD model using a tiny amount of speaker-specific adaptation data for DNN-based speech enhancement. This technique adapts the model by forcing the conditional target distribution (CTD) [23] estimated from the SD model to be close to that estimated from the SI model. The constraint is realized by adding KLD regularization to the ML optimization criterion.

2. Prior Art: MMSE and ML Approaches

2.1. Conventional MMSE criterion

As shown in [12], the prediction error between the target and output could be defined as

$$e_{n,d} = x_{n,d} - \hat{x}_{n,d}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}), \qquad (1)$$

http://dx.doi.org/10.21437/Interspeech.2019-2426

where $\boldsymbol{W}$ is the DNN parameter set to be learned, $\hat{x}_{n,d}$ and $x_{n,d}$ denote the d-th dimension of output and target feature at sample index n respectively, $\boldsymbol{y}_{n-\tau}^{n+\tau}$ is the input feature vector with an acoustic context of $2\tau + 1$. For DNN-based speech enhancement optimized by the conventional MMSE criterion, a mini-batch stochastic gradient descent algorithm is used to minimize the following mean squared error

$$E_{\text{MSE}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}(x_{n,d} - \hat{x}_{n,d}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}))^2, \qquad (2)$$

where D is the size of the target feature vector and N is the mini-batch size.

2.2. ML optimization criterion

In [21], an ML approach to DNN parameter learning by characterizing the prediction error vector as a multivariate Gaussian density with a zero mean vector and an unknown covariance matrix is presented, where the covariance matrix is forced to a diagonal matrix to avoid the instability of the optimization process caused by calculating the inverse of it. This implies that the assumption is made that the prediction error distribution in each dimension independently follows a univariate Gaussian density with a zero mean and an unrestricted variance. Accordingly, we re-derive the ML optimization criterion below. The prediction error shown in Eq. (1) is characterized as a univariate Gaussian density with a zero mean and an unknown variance σ^2:

$$p(e_{n,d}|\sigma_d) = \mathcal{N}(e_{n,d}|0, \sigma_d^2) = \frac{\exp(-\frac{e_{n,d}^2}{2\sigma_d^2})}{\sqrt{2\pi\sigma_d^2}}. \qquad (3)$$

If the target is also a random variable, then the CTD as a function of the input is derived:

$$p(x_{n,d}|\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}, \sigma_d) = \mathcal{N}(x_{n,d}|\hat{x}_{n,d}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}), \sigma_d^2). \quad (4)$$

We assume that the CTDs in all dimensions are independently and identically distributed. Hence we can get the joint CTDs for all dimensions at sample index n:

$$p(\boldsymbol{x}_n|\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}, \boldsymbol{\Sigma}) = \prod_{d=1}^{D}\mathcal{N}(x_{n,d}|\hat{x}_{n,d}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}), \sigma_d^2), \qquad (5)$$

where $\boldsymbol{\Sigma} = \{\sigma_d^2|d = 1, 2, ..., D\}$. Given a training set with N data pairs $(\boldsymbol{Y}, \boldsymbol{X}) = \{(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{x}_n)|n = 1, 2, ...N\}$ and assuming that they are drawn independently from the distribution in Eq. (5), we can define the log-likelihood function as:

$$\ln p(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\Sigma}) = \ln\prod_{n=1}^{N}\prod_{d=1}^{D}\mathcal{N}(x_{n,d}|\hat{x}_{n,d}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}), \sigma_d^2). \qquad (6)$$

Accordingly, the objective function under the probabilistic learning framework is to maximize Eq. (6), which is equivalent to minimizing the following function:

$$E(\boldsymbol{W}, \boldsymbol{\Sigma}) = N\sum_{d=1}^{D}\ln\sigma_d + \sum_{n=1}^{N}\sum_{d=1}^{D}\frac{(x_{n,d} - \hat{x}_{n,d}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}))^2}{2\sigma_d^2}. \qquad (7)$$

Note that the ML criterion shown in Eq. (7) is regressed to the conventional MMSE criterion shown in Eq. (2) when making a strong assumption that the CTD in each dimension independently follows a univariate Gaussian distribution with equal variances.

3. KLD Regularized Adaptation

Experiments using a multi-condition training set built by a large collection of clean data and noise types in [21] have demonstrated the effectiveness of the objective function shown in Eq. (7). Accordingly, the SI model in this paper was optimized by the ML optimization criterion in [21], namely Eq. (7). Based on above-mentioned probabilistic learning framework, we propose a KLD regularized adaptation technique to do adaptation conservatively. The intuition behind this technique is that the CTD estimated from the adapted model should not deviate too far away from that estimated from the unadapted model, especially when the amount of adaptation set is tiny. By adding the KLD as a regularization term to the ML optimization criterion shown in Eq. (7) and removing the terms unrelated to the parameter set $(\boldsymbol{W}, \boldsymbol{\Sigma})$ we can get the regularized optimization criterion

$$E_r(\boldsymbol{W}, \boldsymbol{\Sigma}) = (1 - \rho)E(\boldsymbol{W}, \boldsymbol{\Sigma}) - \rho\sum_{n=1}^{N}\sum_{d=1}^{D}p(x_{n,d}|\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}^{\text{SI}}, \sigma_d^{\text{SI}})\ln p(x_{n,d}|\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}, \sigma_d), \qquad (8)$$

where $p(x_{n,d}|\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}^{\text{SI}}, \sigma_d^{\text{SI}})$ is the CTD from the SI model and computed using the pre-optimized parameters $\boldsymbol{W}^{\text{SI}}$ and $\boldsymbol{\Sigma}^{\text{SI}}$, and ρ is the regularization weight which takes a value between 0 and 1. Taking Eq. (7) into Eq. (8), Eq. (8) can be reorganized to

$$E_r(\boldsymbol{W}, \boldsymbol{\Sigma}) = N(1 - \rho)\sum_{d=1}^{D}\ln\sigma_d + \sum_{n=1}^{N}\sum_{d=1}^{D}\left(\rho p_{n,d}^{\text{SI}}\ln\sigma_d + \frac{(1 - \rho + \rho p_{n,d}^{\text{SI}})(x_{n,d} - \hat{x}_{n,d})^2}{2\sigma_d^2}\right), \qquad (9)$$

where $p_{n,d}^{\text{SI}}$ is a shorthand notation of $p(x_{n,d}|\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}^{\text{SI}}, \sigma_d^{\text{SI}})$ and $\hat{x}_{n,d}$ is a shorthand notation of $\hat{x}_{n,d}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W})$.

An alternating two-step optimization scheme in mini-batch mode is used to optimize $\boldsymbol{W}$ and $\boldsymbol{\Sigma}$ in Eq. (9). First, a closed-form solution of $\boldsymbol{\Sigma}$ can be derived by fixing $\boldsymbol{W}$ and minimizing $E_r(\boldsymbol{W}, \boldsymbol{\Sigma})$ in Eq. (9):

$$\sigma_d = \left(\frac{\sum_{n=1}^{N}(1 - \rho + \rho p_{n,d}^{\text{SI}})(x_{n,d} - \hat{x}_{n,d})^2}{N(1 - \rho) + \sum_{n=1}^{N}\rho p_{n,d}^{\text{SI}}}\right)^{\frac{1}{2}}. \qquad (10)$$

Second, by fixing $\boldsymbol{\alpha}$, $\boldsymbol{W}$ can be optimized by minimizing the following expression:

$$\mathcal{L}(\boldsymbol{W}) = \sum_{n=1}^{N}\sum_{d=1}^{D}\frac{(1 - \rho + \rho p_{n,d}^{\text{SI}})(x_{n,d} - \hat{x}_{n,d})^2}{2\sigma_d^2}. \qquad (11)$$

The back-propagation procedure is used to optimize $\boldsymbol{W}$. The gradient of $\boldsymbol{W}$ is usually obtained by using the chain rule, where only the gradient of the objective function with respect to the DNN output needs to be modified accordingly as shown in Eq. (12), whereas all other derivatives are unaffected.

$$\frac{\partial\mathcal{L}(\boldsymbol{W})}{\partial\hat{x}_{n,d}} = \frac{1}{\sigma_d^2}(1 - \rho + \rho p_{n,d}^{\text{SI}})(\hat{x}_{n,d} - x_{n,d}). \qquad (12)$$

Table 1: *Duration statistics of the utterances of each WSJ0 test speaker for constructing the training and test sets of SD models.*

Speaker ID	440	441	442	443	444	445	446	447
Gender	Male	Female	Female	Male	Female	Female	Male	Male
Total duration (s)	711	694	652	692	675	567	676	673

Note that when the regularization weight ρ is set to 0, the regularized criterion shown in Eq. (9) is regressed to the ML criterion shown in Eq. (7) and the alternating two-step optimization procedure derived above corresponds to those used in [21].

4. Experiments

4.1. Experimental conditions

The 115 noise types which included 100 noise types [24] and 15 home-made noise types were adopted for training to improve the robustness to the unseen noise types. Clean speech utterances were derived from the WSJ0 corpus [25]. The 7138 utterances from 83 speakers denoted as the SI-84 training set were corrupted with the above-mentioned 115 noise types at six levels of SNRs (-5dB, 0dB, 5dB, 10dB, 15dB and 20dB) to build a 12-hour training set for the SI model, consisting of pairs of clean and noisy speech utterances. To make our conclusion more reliable, adaptation experiments were conducted on 8 speakers from the standard Nov92 5K evaluation set, whose detailed duration statistics are given in Table 1. In making the training sets for the 8 SD models, a tiny amount of clean speech utterances from each speaker were fully corrupted with the above-mentioned 115 noise types at above-mentioned six levels of SNRs to build a multi-condition training set for each SD model. The remaining clean speech utterances from each speaker were used to construct the test set for each combination of 15 unseen noise types and SNR levels (-6dB, 0dB, 6dB) for each SD model. The 15 unseen noises [1] from the NOISEX-92 corpus [26] were adopted for testing in this study.

Experiments were conducted on waveforms with 16kHz. The corresponding frame length was set to 512 samples (32msec) with a frame shift of 256 samples. A short-time Fourier transform was used to compute the spectra of each overlapping windowed frame. Then, the 257-dimensional log-power spectra (LPS) features normalized by global mean and variance were employed as the inputs and outputs of the DNNs. In this paper, the feed-forward DNNs were adopted because more powerful DNNs such as recurrent neural networks more easily cause over-fitting for the SD models when the speaker data is little. The network configurations were fixed at three hidden layers, 2048 units for each hidden layer, and 7-frame input ($\tau = 3$). The learning rate for the supervised fine-tuning was set to 0.1 for the first 10 epochs and declined at a rate of 90% after every epoch in the next 40 epochs with the mini-batch size of 128 ($N = 128$). Original phase of noisy speech was adopted with the enhanced LPS for the waveform reconstruction. Short-term objective intelligibility (STOI) [27] and the perceptual evaluation of speech quality (PESQ) [28] were employed to assess the speech intelligibility and quality of processed speech respectively.

Because [21] has demonstrated the effectiveness of the ML optimization criterion shown in Eq. (7) on the multi-condition training set built by a large amount of clean data, the SI model was optimized by the ML optimization criterion rather than the conventional MMSE criterion. Moreover, all the SD models were initialized using the well-trained SI model because the well initialized model can alleviate over-fitting problems with little adaptation data. In addition, the transfer learning approach in [29] is adopted for training all the SD models to further alleviate the over-fitting and mismatch problems, where the strategy that updating the parameters of top 2 layers has been demonstrated to be optimal when the adaptation data is little and thus is adopted here. We select one of the 8 speakers, namely "440" to investigate the experimental details. Finally, we demonstrate the effectiveness of our proposed KLD regularized adaptation technique on all the remaining 7 speakers.

4.2. Experimental results and analysis

Table 3 compares the PESQ and STOI of the SD models optimized by the three objective functions, namely Eq. (2), Eq. (7) and Eq. (9). Here, the regularization weight ρ in the KLD regularized optimization criterion shown in Eq. (9) is set to 1, 1 and 0.7 for 10s, 30s and 68s of adaptation data respectively. Note that the KLD regularized optimization criterion is equivalent to minimizing the KLD of the CTDs from the SI model and the SD model when ρ is set to 1. From this table we can make three observations. First, the conventional MMSE criterion outperforms the ML optimization criterion proposed in [21] when the clean data to build the training set is less than 1 minute. After increasing the clean data to 68s, contrary conclusion is obtained that the ML optimization criterion achieves better objective perceptual quality over the MMSE criterion just as the conclusion drawn in [21]. These imply that the ML optimization criterion is more suitable for the case where clean data for training is enough because it is easier to lead to over-fitting when the amount of clean data is tiny. Second, by comparing the perceptual performance of the SI model shown in Table 2 and SD models optimized by the MMSE criterion using different sizes of adaptation data, we observe that the SD model trained using only 10s adaptation data outperforms the SI model at -6 dB while contrary phenomenon occurs at higher SNRs, especially at 6 dB. After the adaptation data increases to 68s, the SD model achieves significant performance improvements over the SI model at all the SNRs. Third, compared to the MMSE and ML optimization criterion for SD models, our proposed KLD regularized optimization criterion achieves great improvements in STOI especially for extremely little adaptation data and low SNRs. For example, gains of more than 0.04 are achieved in STOI for the 10s of adaptation data at -6dB. Furthermore, the lower the SNR is or the less the adaptation data is, the larger improvements in STOI are obtained. Besides, the KLD adaptation technique yields slight but almost consistent improvements in PESQ with only one exception for the 10s of adaptation data at 6dB. Furthermore, the lower the SNR is, the larger improvements in PESQ are obtained. For example, gains of 0.07 are obtained in PESQ at -6dB while gains of 0.04

[1] N1-N15 noise types: Jet cockpit 1, Jet cockpit 2, Destroyer engine, Destroyer operations, F-16 cockpit, Factory 1, Factory 2, HF channel, Military vehicle, M109 tank, Machine gun, Pink, Volvo, Speech babble and White noise

Table 2: *Average PESQ and STOI of the SI model on the test set of speaker “440” across the 15 unseen noise types.*

-6 dB		0 dB		6 dB	
PESQ	STOI	PESQ	STOI	PESQ	STOI
1.629	0.605	2.100	0.765	2.554	0.880

Table 3: *Average PESQ and STOI comparison for SD models on the test set across the 15 unseen noise types.*

Input SNR: -6 dB						
	10s		30s		68s	
Obj	PESQ	STOI	PESQ	STOI	PESQ	STOI
MMSE	1.651	0.611	1.629	0.633	1.665	0.638
ML	1.638	0.613	1.642	0.631	1.694	0.649
ML+KLD	1.720	0.658	1.699	0.651	1.712	0.655
Input SNR: 0 dB						
	10s		30s		68s	
Obj	PESQ	STOI	PESQ	STOI	PESQ	STOI
MMSE	2.063	0.761	2.123	0.784	2.168	0.790
ML	1.998	0.750	2.114	0.777	2.180	0.794
ML+KLD	2.095	0.794	2.177	0.804	2.215	0.803
Input SNR: 6 dB						
	10s		30s		68s	
Obj	PESQ	STOI	PESQ	STOI	PESQ	STOI
MMSE	2.436	0.858	2.545	0.881	2.615	0.888
ML	2.287	0.834	2.500	0.868	2.604	0.885
ML+KLD	2.394	0.872	2.581	0.898	2.651	0.897

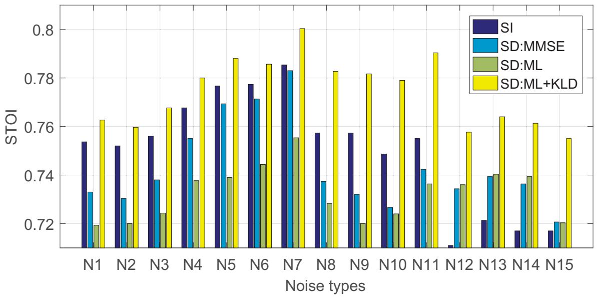

Figure 1: *Averaged STOI of the SD models using 30s of data and the SI model on the test set across all the SNRs .*

are achieved at 6dB for the 30s of adaptation data. Figure 1 presents the effects of different noise scenarios, where it can be observed that our proposed KLD regularized optimization criterion achieves best speech intelligibility performance for all SNR levels and 15 unseen noise types. Moreover, Figure 2 shows a spectrogram comparison, where we observe that the SD models perform better in noise reduction over the SI model and the proposed KLD regularized optimization criterion achieves less speech distortions.

Figure 3 presents average performance comparison among SI model and SD models for the other 7 speakers, where 30s of clean data is adopted for each speaker. It can be observed that the KLD regularized optimization criterion achieves consistent improvements in STOI and PESQ over the MMSE and ML optimization criterion for each SD model. Moreover, compared to the SI model, it yields consistent much better STOI and a little better PESQ with only one exception for the speaker “443”.

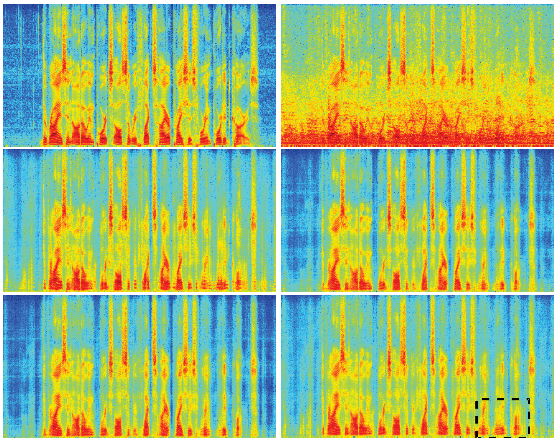

Figure 2: *Spectrograms of an utterance tested with Speech babble noise at SNR=0 dB for 30s of adaptation data (from left to right and from up to down): clean speech, noisy speech, SI, SD:MMSE, SD:ML, SD:ML+KLD.*

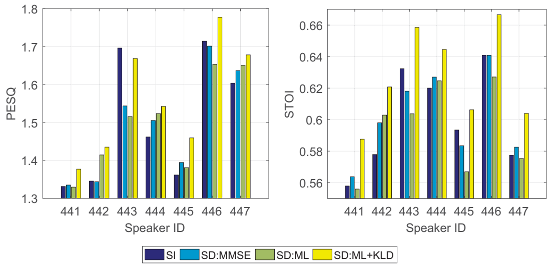

Figure 3: *Averaged PESQ and STOI comparison across the 15 unseen noise types for the remaining seven speakers on their respective test sets at -6dB.*

5. Conclusion

Based on our previously developed maximum likelihood learning framework for parameter optimization in DNN-based speech enhancement, we propose a KLD regularized adaptation approach to adapting a well-trained SI model to another speaker-specific SD model using a tiny amount of adaptation data. The regularized optimization criterion is derived, by adding to the ML optimization criterion the KLD between the conditional target densities estimated from the SI and SD model. Experiments demonstrate that the proposed KLD framework achieves consistent speech intelligibility performance improvements over the MMSE and ML optimization criteria under all the SNR levels and 15 unseen noise types evaluated for all eight tested speakers. In future, we will combine it with other advanced transfer learning techniques to achieve much better performance of the SD model over the SI model using much less adaptation data.

6. 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, and Tencent.

7. References

[1] 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.

[2] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’79.*, vol. 4. IEEE, 1979, pp. 208–211.

[3] K. Paliwal, K. Wójcicki, and B. Schwerin, “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain,” *Speech communication*, vol. 52, no. 5, pp. 450–475, 2010.

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

[5] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[6] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[7] T. Gerkmann and R. C. Hendriks, “Unbiased mmse-based noise power estimation with low complexity and low tracking delay,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[8] 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.

[9] P. C. Loizou, *Speech Enhancement: Theory and Practice.* CRC Press, 2013.

[10] P. C. Loizou and G. Kim, “Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions,” *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 47–56, 2011.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.

[12] ——, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

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

[14] Y. Wang, A. Narayanan, and D. 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.

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

[16] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 483–492, 2016.

[17] T. Gao, J. Du, L. Xu, C. Liu, L.-R. Dai, and C.-H. Lee, “A unified speaker-dependent speech separation and enhancement system based on deep neural networks,” in *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP).* IEEE, 2015, pp. 687–691.

[18] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, “A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments,” *Speech Communication*, vol. 95, pp. 28–39, 2017.

[19] Q. Wang, S. Wang, F. Ge, C. W. Han, J. Lee, L. Guo, and C.-H. Lee, “Two-stage enhancement of noisy and reverberant microphone array speech for automatic speech recognition systems trained with only clean speech,” in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP).* IEEE, 2018, pp. 21–25.

[20] L. Sun, J. Du, T. Gao, Y. Fang, F. Ma, J. Pan, and C.-H. Lee, “A two-stage single-channel speaker-dependent speech separation approach for chime-5 challenge,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 6650–6654.

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

[22] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, “KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 2013, pp. 7893–7897.

[23] D. A. Nix and A. S. Weigend, “Learning local error bars for nonlinear regression,” in *Advances in neural information processing systems*, 1995, pp. 489–496.

[24] G. Hu, “100 nonspeech environmental sounds,[online] available: http://web. cse. ohio-state. edu/pnl/corpus/hunonspeech,” *HuCorpus. html*, 2004.

[25] J. Garofalo, D. Graff, D. Paul, and D. Pallett, “Csr-i (wsj0) complete,” *Linguistic Data Consortium, Philadelphia*, 2007.

[26] 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.

[27] 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.

[28] L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida, “Perceptual evaluation of blind source separation for robust speech recognition,” *Signal Processing*, vol. 88, no. 10, pp. 2578–2583, 2008.

[29] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “Cross-language transfer learning for deep neural network based speech enhancement,” in *The 9th International Symposium on Chinese Spoken Language Processing.* IEEE, 2014, pp. 336–340.