INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A Maximum Likelihood Approach to Deep Neural Network Based Nonlinear Spectral Mapping for Single-Channel Speech Separation

Yannan Wang[1], *Jun Du*[1], *Li-Rong Dai*[1], *and 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

wyn314@mail.ustc.edu.cn, jundu@ustc.edu.cn, lrdai@ustc.edu.cn, chl@ece.gatech.edu

Abstract

In contrast to the conventional minimum mean squared error (MMSE) training criterion for nonlinear spectral mapping based on deep neural networks (DNNs), we propose a probabilistic learning framework to estimate the DNN parameters for single-channel speech separation. A statistical analysis of the prediction error vector at the DNN output reveals that it follows a unimodal density for each log power spectral component. By characterizing the prediction error vector as a multivariate Gaussian density with zero mean vector and an unknown covariance matrix, we present a maximum likelihood (ML) approach to DNN parameter learning. Our experiments on the Speech Separation Challenge (SSC) corpus show that the proposed learning approach can achieve a better generalization capability and a faster convergence than MMSE-based DNN learning. Furthermore, we demonstrate that the ML-trained DNN consistently outperforms MMSE-trained DNN in all the objective measures of speech quality and intelligibility in single-channel speech separation.

Index Terms: the prediction error, Gaussian density, maximum likelihood estimation, deep neural network, speech separation

1. Introduction

Speech separation [1] is the task of separating a speech component of interest from speech segments mixed with multiple speakers while single-channel speech separation refers to the more challenging situation when only one microphone is available for recording. It is widely used in many major real-work applications, e.g., automatic speech recognition (ASR) [2] in the Speech Separation Challenge (SSC) [3].

Many signal processing methods to recover clean speech by estimating the ideal Wiener filter have been proposed in the context of minimum mean squared error (MMSE) [4]. Moreover, computational auditory scene analysis (CASA) [5], inspired by the ability of human auditory perception to revert signals of interest from background distractions, is widely adopted without assuming any knowledge about mixed speakers in an unsupervised separation mode. For instance, a tandem algorithm to perform pitch estimation and segregate voiced portions of target speech jointly and iteratively is proposed in [6]. Onset/offset-based segmentation and model-based grouping are introduced to manage unvoiced portions in [7]. Unsupervised clustering for sequential grouping is adopted in [8] by maximizing the ratio of between-cluster and within-cluster distances while penalizing within-cluster concurrent pitches. Recently, a data-driven approach to separate the underlying clean speech segments by matching each mixed speech segment against a composite training segment is presented in [9]. Another popular approach is non-negative matrix factorization (NMF) [10, 11, 12] and probabilistic latent semantic indexing (PLSI) [13, 14] which factorize time-frequency spectral representations by decomposing speech signal into sets of bases and weight matrices.

On the other hand, model-based approaches are widely used in a supervised mode which generally builds speaker-dependent models with known identities of mixed speakers. For example, Roweis [15] employs the factorial hidden Markov model (FHMM) to learn the information of a speaker and then separate the speech mixture through computing a mask function and refiltering. Factorial-max vector quantization (MAXVQ) is introduced as a probabilistic model in [16]. The layered factorial HMM, incorporating temporal and grammar dynamics [17], performs quite well in speech separation and the Gaussian mixture model (GMM) is used in [18, 19] to re-synthesize the speech signals. An iterative GMM-based approach based on a maximum a posteriori (MAP) estimator to overcome possible mismatches between the training and test conditions improves separation results significantly in [20].

Deep learning methods have been explored in speech enhancement and ideal binary mask estimation in some recent work [21, 22]. And then a discriminative training objective is proposed in [23] which takes into account the similarity between the prediction and other sources when minimizing the squared error between the output of neural network and the target reference. Another generic discriminative training criterion corresponding to optimal source reconstruction from time-frequency masks is also validated in a reduced feature space [24]. Besides, weighted denoising auto-encoder is studied in [25] which emphasizes different frequency bands empirically. Furthermore, in [26] the perceptual weighting deep neural network (DNN) uses perceptual weighting matrix to adjust the weight of the prediction error. In this study, different from the conventional approaches to improve the objective function design of DNN based on the MMSE framework (MMSE-DNN), we explore the maximum likelihood (ML) solution within the probabilistic learning framework to optimize DNN parameters with the assumption that the prediction error vector of the regression DNN follows a multivariate Gaussian density. Accordingly, a training procedure of ML-trained DNN (ML-DNN) is designed to update both DNN parameters and the covariance matrix of Gaussian density alternatively. The MMSE-DNN approach could be considered as a special case of the proposed ML-DNN approach with an identity covariance matrix. The evaluation on the SSC corpus show that the proposed ML-DNN approach achieves a significantly better separation performance than the conventional MMSE-DNN approach. Moreover, the ML-DNN approach can also yield a better generalization capability and a faster convergence.

http://dx.doi.org/10.21437/Interspeech.2017-830

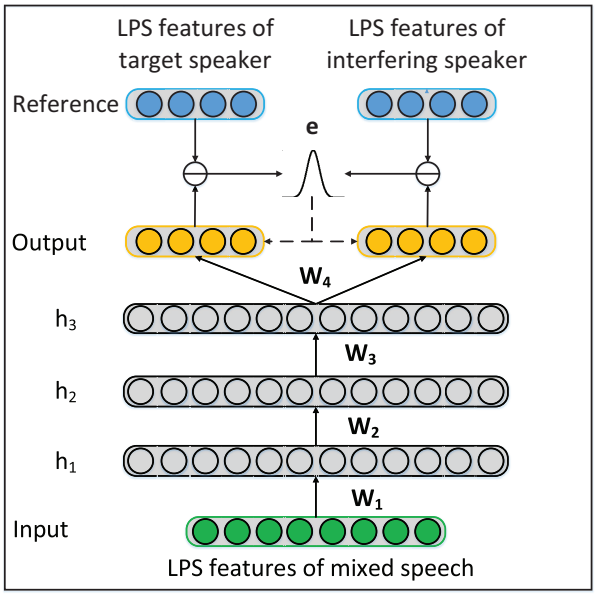


Figure 1: *The ML-DNN architecture for speech separation.*

2. The Proposed ML-DNN Approach

DNN is essentially a feed-forward multi-layer perceptron with many hidden layers [27, 28]. Recently it has been widely used for classification tasks in image and speech areas [29, 30]. In our recent work for speech enhancement [31], DNN was instead adopted as a regression model to learn the relationship between noisy and clean speech under the conventional MMSE training criterion. Furthermore, in [32, 33, 34] DNN is adopted to model the highly non-linear mapping relationship from mixed speech to the target and interfering signals in a supervised or semi-supervised mode.

In this study, to further improve the generalization capability of the conventional MMSE optimization for the regression DNN, we redefine the objective function in the probabilistic framework and adopt the maximum likelihood estimation for the parameter learning, as shown in Figure 1. The input of DNN is the $(2\tau+1)D$-dimensional log-power spectral (LPS) feature vector of mixed speech with an acoustic context of $2\tau+1$ neighbouring frames while the dual output refers to a $2D$-dimensional concatenation of two LPS feature vectors corresponding to the target speaker and the interfering speaker. The sigmoidal hidden units and linear output units are adopted. Suppose the DNN dual output vector is $\hat{\boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{W})$ with the input vector $\boldsymbol{x}$ and the DNN parameter set $\boldsymbol{W}$ while the corresponding reference vector is $\boldsymbol{y}$. The prediction error vector $\boldsymbol{e}$ could be defined as:

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{W}) \tag{1}$$

which is assumed to follow a multivariate Gaussian density with a $2D$-dimensional zero mean vector and a $2D \times 2D$ covariance matrix $\boldsymbol{\Sigma}$:

$$p(\boldsymbol{e}) = \mathcal{N}(\boldsymbol{e}; \boldsymbol{0}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^D |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{e}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{e}\right) \tag{2}$$

If the reference vector $\boldsymbol{y}$ is also a random vector, then Eq. (2) is equivalent to:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{y}; \hat{\boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{W}), \boldsymbol{\Sigma}) \tag{3}$$

which implies that the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ with the parameter set $(\boldsymbol{W}, \boldsymbol{\Sigma})$ is unimodal. Given a training set with N data pairs $(\boldsymbol{X}, \boldsymbol{Y}) = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)|n = 1, 2, ...N\}$ and making the assumption that these data pairs are drawn independently from the distribution in Eq. (3), we can define the likelihood function as:

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n; \hat{\boldsymbol{y}}_n(\boldsymbol{x}_n, \boldsymbol{W}), \boldsymbol{\Sigma}) \tag{4}$$

where the parameter set $(\boldsymbol{W}, \boldsymbol{\Sigma})$ is to be optimized. Accordingly, the log-likelihood function can be written as:

$$\ln p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \mathcal{N}(\boldsymbol{y}_n; \hat{\boldsymbol{y}}_n(\boldsymbol{x}_n, \boldsymbol{W}), \boldsymbol{\Sigma}) = C -$$

$$\frac{N}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n(\boldsymbol{x}_n, \boldsymbol{W}))^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n(\boldsymbol{x}_n, \boldsymbol{W})) \tag{5}$$

where C is a constant. We adopt maximum likelihood criterion to alternatively optimize $\boldsymbol{W}$ and $\boldsymbol{\Sigma}$. To maximize Eq. (5) with respect to $\boldsymbol{W}$, it is equivalent to minimizing the following sum-of-squares error function in terms of Mahalanobis distance:

$$E(\boldsymbol{W}) = \sum_{n=1}^{N} (\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n(\boldsymbol{x}_n, \boldsymbol{W}))^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n(\boldsymbol{x}_n, \boldsymbol{W})). \tag{6}$$

Then the back-propagation procedure with a stochastic gradient descent method is used to optimize $\boldsymbol{W}$ in the mini-batch mode of M sample frames.

Alternatively, we can also maximize Eq. (5) with respect to $\boldsymbol{\Sigma}$. Then the update formula can be derived as:

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n(\boldsymbol{x}_n, \boldsymbol{W}))(\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n(\boldsymbol{x}_n, \boldsymbol{W}))^\top \tag{7}$$

To avoid the instability of the optimization process by calculating the inverse of the covariance matrix $\boldsymbol{\Sigma}$, we use the diagonal covariance matrix in this study. The whole training procedure is summarized as follows.

Algorithm 1 Procedure of ML-DNN training

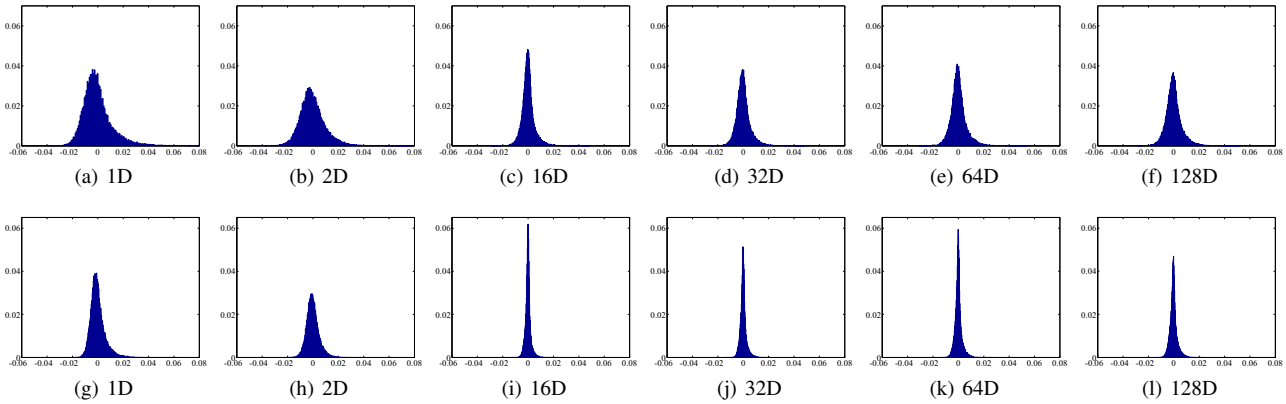
Step 1: Initialization
Initialize the DNN parameter set $\boldsymbol{W}$ randomly. The covariance matrix $\boldsymbol{\Sigma}$ is set to an identity matrix.

Step 2: Fix $\boldsymbol{\Sigma}$ and update $\boldsymbol{W}$
By minimizing Eq. (6) with N training sample pairs, the back-propagation procedure with a stochastic gradient descent method is used to update $\boldsymbol{W}$ in the mini-batch mode of M sample frames.

Step 3: Fix $\boldsymbol{W}$ and update $\boldsymbol{\Sigma}$
Update $\boldsymbol{\Sigma}$ via Eq. (7).

Step 4: Go to Step 2 for L epochs

We should indicate that the conventional MMSE-DNN is a special case of ML-DNN where the covariance matrix $\boldsymbol{\Sigma}$ in Eq. (5) is always an identity matrix, namely making a strong assumption that all the LPS components are with equal variances. This is the reason why MMSE optimization often leads to a poor generalization capability.



(a) 1D (b) 2D (c) 16D (d) 32D (e) 64D (f) 128D

(g) 1D (h) 2D (i) 16D (j) 32D (k) 64D (l) 128D

Figure 2: *The distributions for selected dimensions of the prediction error vector from well-trained DNN on the cross validation set: (a)-(f) refer to MMSE-DNN while (g)-(l) correspond to ML-DNN.*

3. Statistical Analysis on Prediction Errors

For the derivation of ML solution in Section 2 , we assume that the prediction error vector $\boldsymbol{e}$ follows a multivariate Gaussian density with zero mean. To verify the reasonability of this assumption, we first explore the distribution of the prediction error of regression DNN for speech separation. We present the distributions of selected dimensions (1, 2, 16, 32, 64, 128) of the prediction error vector which is the LPS feature on the cross validation set for both well-trained MMSE-DNN and ML-DNN as shown in Figure 2 which tells the percentage of different range of values. It is observed that all selected dimensions of the prediction error vector approximately follows a unimodal distribution with the mean closing to zero for both MMSE-DNN and ML-DNN. However the variance of each dimension is quite different, which implies that the assumption of equivalent variances in MMSE-DNN is not reasonable. Actually this observation is also the main motivation of this study, namely adopting a multivariate Gaussian density with a zero mean vector and an unrestricted covariance matrix to represent the prediction error vector and employing the maximum likelihood criterion to optimize both the DNN parameters and the covariance matrix. Furthermore, for all selected dimensions, the mean of each dimension in ML-DNN is closer to zero than MMSE-DNN while the variance of each dimension in ML-DNN is also much smaller than that in MMSE-DNN, demonstrating that ML-DNN could better model the prediction errors.

Besides, we also compare the generalization capability between MMSE-DNN and ML-DNN via the learning curves of the reconstruction loss on the cross validation set, as illustrated in Figure 3. The reconstruction loss refers to the mean squared error adopted as the learning objective in MMSE-DNN. Interestingly, we observe that the MMSE-DNN to minimize the reconstruction loss on the training data consistently generates larger errors on the cross validation set than ML-DNN which is maximizing the likelihood rather than directly minimizing the reconstruction loss in the training stage. So it is clear that ML-DNN can achieve a better generalization capability than MMSE-DNN. Moreover, the learning curve of ML-DNN shows a faster convergence than MMSE-DNN. Actually, based on Figure 2, the smaller variances of ML-DNN can also easily deduce the smaller reconstruction loss of ML-DNN in Figure 3, as the reconstruction loss is the summation of variances across all dimensions.

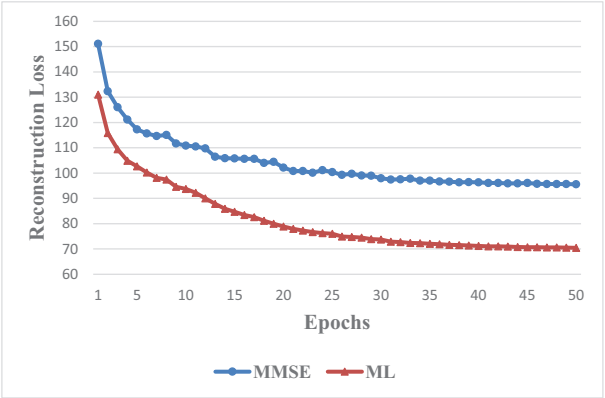


Figure 3: *The learning curve comparison of reconstruction loss between MMSE-DNN and ML-DNN on cross validation set.*

4. Experiments and Results

Our experiments were conducted on the SSC corpus [35] which was generated by mixing target utterances with another simultaneous masker utterance by a competing speaker with a very similar structure [3]. For training of regression DNN, all the utterances of the target speakers were used to generate the mixtures by adding the interfering speech segments to the target speech with signal-to-noise-ratios (SNRs) ranging from -10dB to 10dB using an increment of 2dB. This gave a good coverage of SNRs in the test set whose SNR level goes from -9dB to 6dB with an increment of 3dB. The evaluation was conducted in two modes, namely supervised mode and semi-supervised mode. In the supervised mode, both the training data of one target speaker and one interfering speaker can be provided for building the regression DNN with dual outputs corresponding to these speakers. In the semi-supervised mode, only the training data of the target speaker is given and the interfering speaker is unseen. Therefore, the training data of multiple other speakers should be used to simulate the unseen interfering speaker.

We down-sampled the original waveforms from 25kHz to 16kHz. The frame length and shift are 512 samples (32 msec) and 256 samples (16 msec), respectively. A short-time Fourier transform was adopted to compute the spectra of each overlapping windowed frame. Then 257-dimensional (D=257) log-

Table 1: *Average performance comparison of different objective measures between MMSE-DNN and ML-DNN in supervised mode.*

Input SNR (dB)		-9	-6	-3	0	3	6
Output SNR	MMSE	6.20	7.30	8.47	9.74	11.05	12.34
	ML	6.92	8.20	9.62	11.14	12.72	14.29
STOI	MMSE	1.06	1.09	1.12	1.14	1.16	1.17
	ML	1.09	1.12	1.14	1.16	1.18	1.19
PESQ	MMSE	3.10	3.33	3.54	3.74	3.90	4.05
	ML	3.29	3.53	3.73	3.91	4.08	4.24
SDR	MMSE	4.03	5.64	7.24	8.88	10.49	12.05
	ML	5.11	6.82	8.59	10.44	12.27	14.06
SIR	MMSE	14.94	16.23	17.59	19.13	20.83	22.74
	ML	15.37	16.55	17.93	19.56	21.35	23.30
SAR	MMSE	5.49	7.07	8.65	10.25	11.79	13.24
	ML	6.56	8.31	10.11	11.95	13.76	15.48

power spectral features were used to train DNNs. For the waveform reconstruction, the original phase of mixed speech was adopted with the separated log-power spectra [36]. In all experiments the DNN consisted of 1799 input nodes (τ=3), 3 hidden layers with 2048 sigmoidal nodes per layer, and 514 dual output nodes. For the update of DNN parameters in both MMSE-DNN and ML-DNN, the learning rate for the supervised fine-tuning was set to 0.1 for the first 10 epoch and declined at a rate of 90% after every epoch in the next 40 epochs (L=50) with the mini-batch size of M=256. The global normalization was also applied to the input features to guarantee the zero mean and unit variance.

4.1. Evaluation in supervised mode

For evaluation in the supervised mode, 8 combinations of targets and interferers were selected to generate the training and evaluation set. And they were equally assigned for four possible gender combinations, namely female and female (F + F), male and male (M + M), female and male (F + M), male and female (M + F). About 50 hours of mixed speech were synthesized for training the corresponding DNNs for each combination. Output SNR for measuring noise reduction, STOI for measuring speech intelligibility [37] and PESQ for measuring speech quality [38] are compared in Table 1. Clearly, the proposed ML-DNN approach yielded consistent and significant improvements over the conventional MMSE-DNN approach for all input SNR levels, e.g., an output SNR gain of 1.15dB at the input SNR of -3dB, STOI rising from 1.09 to 1.12 at the input SNR of -6 dB, and about 0.2 PESQ gain for all input SNR levels.

We also examined the effectiveness of ML-DNN with three measures designed for speech separation: Source to Interference Ratio (SIR), Source to Artifacts Ratio (SAR), and Source to Distortion Ratio (SDR) shown in Table 1, according to the BSS-EVAL metrics [39]. We can observe that the gain of ML-DNN over MMSE-DNN is between 1dB and 2dB for both SDR and SAR across all input SNR levels, e.g, SDR rising from 8.88dB to 10.44dB at the input SNR of 0dB, which indicated that the less artifacts were introduced by the proposed ML-DNN to separate the mixed speech. Moreover, the interference was also suppressed more significantly by ML-DNN as SIR increased about 0.5dB in average.

Table 2: *Average performance comparison of different objective measures between MMSE-DNN and ML-DNN in semi-supervised mode.*

Input SNR (dB)		-9	-6	-3	0	3	6
Output SNR	MMSE	3.43	4.26	5.20	6.23	7.31	8.32
	ML	3.73	4.64	5.71	6.87	8.05	9.16
STOI	MMSE	0.78	0.82	0.85	0.88	0.91	0.93
	ML	0.80	0.83	0.86	0.89	0.92	0.94
PESQ	MMSE	2.09	2.27	2.43	2.60	2.78	2.95
	ML	2.15	2.33	2.52	2.70	2.89	3.06
SDR	MMSE	0.70	2.11	3.60	5.16	6.74	8.19
	ML	1.27	2.72	4.28	5.90	7.46	8.91
SIR	MMSE	7.09	8.30	9.71	11.33	13.20	15.26
	ML	7.72	9.00	10.49	12.21	14.05	15.98
SAR	MMSE	2.97	4.21	5.52	6.86	8.22	9.42
	ML	3.42	4.71	6.09	7.51	8.90	10.14

4.2. Evaluation in semi-supervised mode

In the semi-supervised mode, one target and 6 interferers were adopted in the training stage to generate the mixed speech with two speakers. In the evaluation stage for separating the target, the interfering speaker is unknown, namely not included in 6 interferers at the training stage. In this study about 50 hours of training data were used to train each of two DNNs with one male target and one female target, respectively. And all the mixtures with those two targets in the test set were evaluated in the following experiments. The separation results are displayed in Table 2 with remarkable improvements from MMSE-DNN to ML-DNN for all measures. For example, STOI increases from 0.91 to 0.92 and PESQ rises from 2.78 to 2.89 at the input SNR of 3dB. Besides, the gain from 0.5dB to 0.8dB was observed for SDR, SIR and SAR under different input SNR levels, e.g. SDR increasing from 5.16dB to 5.90dB at the input SNR of 0dB. Finally, the improvements in semi-supervised mode is smaller than those in supervised mode because the covariance matrix of the prediction error is related to speaker characteristics. The assumption of the unimodal Gaussian distribution can not be fully satisfied when multiple interfering speakers are introduced.

5. Conclusion and Future Work

In this study we proposed a novel maximum likelihood approach to DNN-based speech separation with a reasonable assumption that the prediction error vector of DNN follows the Gaussian distribution. In the ML solution, both the DNN parameters and the covariance matrix of the prediction error vector are jointly and alternatively optimized. The proposed learning approach can automatically reinvest all frequency bands with different significance and reduce the errors in the propagation process. Compared with the conventional MMSE optimization, our approach could achieve a smaller reconstruction loss and a better generalization capability.

6. Acknowledgment

This work was partially funded by the National Nature Science Foundation of China (Grant No. 61671422 and U1613211), the National Key Research and Development Program of China (Grant No. 2016YFB1001300) and also supported by Tencent.

7. References

[1] V. C. Shields, "Separation of added speech signals by digital comb filtering," *S.M. Thesis, Dept. of Electrical Engineering, MIT*, 1970.

[2] K.-C. Yen and Y. Zhao, "Co-channel speech separation for robust automatic speech recognition: stability and efficiency," in *Proc. ICASSP*, vol. 2, 1997, pp. 859–862.

[3] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Comput. Speech Lang*, vol. 24, no. 1, pp. 1–15, 2010.

[4] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[5] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[6] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[7] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Comput. Speech Lang*, vol. 24, no. 1, pp. 77–93, 2010.

[8] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 122–131, 2013.

[9] J. Ming, R. Srinivasan, D. Crookes, and A. Jafari, "Closea data-driven approach to speech separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1355–1368, 2013.

[10] M. Schmidt and R. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. Interspeech*, 2006, pp. 2614–2617.

[11] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proc. ICASSP*, 2015.

[12] O. Dikmen and A. T. Cemgil, "Unsupervised single-channel source separation using bayesian nmf," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*. IEEE, 2009, pp. 93–96.

[13] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

[14] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," *Advances in models for acoustic processing, NIPS*, vol. 148, pp. 8–1, 2006.

[15] S. T. Roweis, "One microphone source separation," in *NIPS*, vol. 13, 2000, pp. 793–799.

[16] ——, "Factorial models and refiltering for speech separation and denoising," in *Proc. Eurospeech*, 2003, pp. 1009–1012.

[17] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Comput. Speech Lang*, vol. 24, no. 1, pp. 45–66, 2010.

[18] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1766–1776, 2007.

[19] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2299–2310, 2007.

[20] K. Hu and D. Wang, "An iterative model-based approach to cochannel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 14, 2013.

[21] 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 *Proc. Interspeech*, 2012, pp. 22–25.

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

[23] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. ICASSP*, 2014, pp. 1562–1566.

[24] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*. IEEE, 2014, pp. 577–581.

[25] 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.

[26] W. Han, X. Zhang, G. Min, X. Zhou, and W. Zhang, "Perceptual weighting deep neural networks for single-channel speech enhancement," in *Intelligent Control and Automation (WCICA), 2016 12th World Congress on*. IEEE, 2016, pp. 446–450.

[27] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag*, vol. 29, no. 6, pp. 82–97, 2012.

[28] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. ICASSP*, 2013, pp. 8599–8603.

[29] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3642–3649.

[30] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[31] Y. Xu, J. Du, L. 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.

[32] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. ICSP*, 2014, pp. 473–477.

[33] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *Proc. ISCSLP*, 2014, pp. 250–254.

[34] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 24, no. 8, pp. 1424–1437, 2016.

[35] M. Cooke and T.-W. Lee, "Speech separation challenge," [http://staffwww.dcs.shef.ac.uk/people/M.Cooke/Speech SeparationChallenge.htm], 2006.

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

[37] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

[38] ITU-T and R. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs ," *International Telecommunication Union-Telecommunication Standardisation Sector*, 2001.

[39] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.