ALLEVIATING THE LOSS-METRIC MISMATCH IN SUPERVISED SINGLE-CHANNEL SPEECH ENHANCEMENT

Yang Yang[1,2,3], *Hui Zhang*[1,2,3], *Xueliang Zhang*[1,2,3], *Huaiwen Zhang*[1,2,3,*]

[1]College of Computer Science, Inner Mongolia University, China

[2]National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian

[3]Inner Mongolia Key Laboratory of Mongolian Information Processing Technology

yangyang@mail.imu.edu.cn,{cszh, cszxl, huaiwen.zhang}@imu.edu.cn

ABSTRACT

In this paper, we study the loss-metric mismatch problem of supervised single-channel speech enhancement system. Most of the existing speech enhancement systems achieve unsatisfying performance since their empirically selected loss functions have semantic gaps with the non-differentiable evaluation metrics, a.k.a., the loss-metric mismatch problem. In this work, we propose a simple yet efficient method to generate suitable loss functions for the real front-end speech enhancement scenarios to alleviate the loss-metric mismatch problem. Specifically, we adopt the function smoothing technique and approximate the non-differentiable evaluation metrics by a set of basis functions and their linear combination. Experimental results demonstrate that the loss function generated by our method helps the speech enhancement system achieve remarkable performance in most evaluation metrics than the traditional empirically selected ones.

Index Terms— Supervised Single-Channel Speech Enhancement, Loss-Metric Mismatch, Function Smoothing.

1. INTRODUCTION

The single-channel speech enhancement task aims to improve the intelligibility and perceived quality of degraded speech signals. It benefits many downstream applications, including robust automatic speech recognition (ASR), mobile speech communication, and speaker recognition [1]. Early efforts, such as spectrum subtraction [2], Wiener filtering [3], and non-negative matrix factorization [4] achieve good performance with the assumption that the noise in speech enhancement is stationary. With the development of computational auditory scene analysis (CASA) [5], the speech enhancement problem has been considered as a supervised learning [6] problem, in which the non-stationary noise can be well captured.

Although existing supervised single-channel speech enhancement systems achieve state-of-the-art performance, they suffer from the loss-metric mismatch problem [7]. Specifically, the loss function of speech enhancement task is usually the Mean Square Error (MSE), while the evaluation metrics are Short-Time Objective Intelligibility (STOI) [8] and Perceptual Evaluation of Speech Quality (PESQ) [9]. Some studies [10, 11] have demonstrate that the lower MSE score does not guarantee higher quality and clarity. There are semantic gaps between loss functions and evaluation metrics, resulting in inefficient model training since the reduction of loss function does not reflect the improvement of the evaluation metrics.

The most intuitive way is to optimize the evaluation metrics directly. However, it is impossible since the evaluation metrics are non-differentiable. For example, the widely used PESQ [9] is discontinuous. No gradient can be calculated and trained for these metrics. Recently, various sophisticated methods have been proposed to tackle the mismatch problem. For example, [12] adopts reinforcement learning method [13], which uses evaluation metrics as a reward function to optimize the ASR accuracy directly. [14] trains a PESQ neural network to fit the evaluation metrics, and uses the trained network as loss function to optimize the speech enhancement network. [15] proposed a complex method to approximately calculate the STOI and PESQ loss function gradients to improve the STOI and PESQ directly. [16] has focused on STOI score optimization to improve speech intelligibility.

Except these above complicated methods, function smoothing [17, 18] is a more simple and common used method to approximate the non-differentiable function with a differentiable ones. It is usually implemented by replacing the non-differentiable operation with a set of differentiable operations. By using function smoothing, an approximate trainable loss function can be obtained with the computation graph analysis. For instance, the $\max$ operation can be substituted with $\log$, sum and $\exp$ in the computation graph of the original loss function: $\max(x_1, x_2, \ldots, x_n) \approx \log\left(\sum_{i=1}^{n} e^{x_i}\right)$. If the computation graph of the original loss function, i.e., evaluation metrics is difficult to involve,

* Huaiwen Zhang is the corresponding author.

the smoothing version can also be obtained by some fitting methods on sampling data.

In this paper, we propose a simple yet effective method to generate suitable loss functions for supervised speech enhancement system, which perform consistently with evaluation metrics. Unlike the standard function smoothing method, which fits the value of evaluation metrics by differentiable basis functions, we obtain a suitable loss function by restricting the basis function to have the same tendency with evaluation metrics. We first introduce a series of basis functions to approximate the non-differentiable evaluation metrics. Then the correlation coefficients between these basis functions and evaluation metrics are calculated to weighty generate the suitable loss functions for training. Experimental results showed that the loss function of the approximate evaluation metrics selected by our strategy achieves the best performances on most metrics.

2. PROBLEM SETTING

The single-channel speech enhancement task aims to enhance the target speech $C(n)$ from additive noise $N(n)$ with only one single microphone signal $M(n)$:

$$M(n) = C(n) + N(n) \tag{1}$$

where n is the sampling points index.

Most speech enhancement methods transform the $M(n)$, $C(n)$, and $N(n)$ into frequency domain by short-time Fourier transform (STFT). Since the STFT is a linear transformation, the relationship among the signal is maintained:

$$M(t,f) = C(t,f) + N(t,f) \tag{2}$$

where t and f are the time and frequency index, respectively.

Two categories of methods are used in speech enhancement: the mapping-based [19, 20] and the mask-based [21] methods. The mapping-based method estimates $\hat{C}(t,f)$ from $M(t,f)$, directly. The mask-based method estimates a mask $\hat{E}(t,f)$ from $M(t,f)$, then the $\hat{C}(t,f)$ is obtained by multiplied the estimated mask $\hat{E}(t,f)$ and the noisy signal $M(t,f)$.

$$\hat{C}(t,f) = \hat{E}(t,f) * M(t,f) \tag{3}$$

where $*$ is the point-to-point multiplication, and the hat symbol indicates the estimation.

Both the mapping-based and mask-based methods formalize the speech enhancement task as a regression task. Thus, the Mean Square Error (MSE) is adopted as the loss function by experience. For the mapping-based loss function, the MSE is used to measure the difference between the clean speech $C(t,f)$ and its estimation $\hat{C}(t,f)$:

$$MSE = |C(t,f) - \hat{C}(t,f)|^2 \tag{4}$$

For the masked-based loss function, the MSE is used to measure the difference between the ideal mask $E(t,f)$ and its estimation $\hat{E}(t,f)$:

$$MSE = |E(t,f) - \hat{E}(t,f)|^2 \tag{5}$$

The speech enhancement system is usually evaluated with STOI for the intelligibility and PESQ for the quality. However, the MSE loss is not a good indicator, some studies [10, 11] have pointed out that the lower MSE score does not guarantee higher quality and clarity. There is a clear loss-metric mismatch problem.

3. THE PROPOSED METHOD

To alleviate the loss-metric mismatch problem, we propose a function smoothing style method to obtain a suitable loss function for the speech enhancement system. We first define a series of basis functions to approximate the non-differentiable evaluation metrics. Then, the correlation coefficients between these basis functions and evaluation metrics are calculated, which restricting the basis function to have the same tendency with evaluation metrics. With the correlation coefficients, we can weightedly generate suitable loss functions for training.

3.1. The Basis Functions

We define a list of divergence functions as the basis loss functions:

$$KL(x,y) = x \cdot \log \frac{x}{y} \tag{6}$$

$$symKL(x,y) = x \cdot \log \frac{x}{y} + y \cdot \log \frac{y}{x} \tag{7}$$

$$GKL(x,y) = x \cdot \log \frac{x}{y} - (x-y) \tag{8}$$

$$rGKL(x,y) = y \cdot \log \frac{y}{x} - (y-x) \tag{9}$$

$$JS(x,y) = \frac{1}{2}\left(x \cdot \log \frac{2x}{x+y} + y \cdot \log \frac{2y}{x+y}\right) \tag{10}$$

$$IS(x,y) = \frac{x}{y} - \log \frac{x}{y} - 1 \tag{11}$$

$$rIS(x,y) = \frac{y}{x} - \log \frac{y}{x} - 1 \tag{12}$$

where, the Equation 6, 7 , 8 and 9 are the KL, GKL, JS, and JS divergence, respectively. The $rGKL$ and rIS are the reverse version of the GKL and IS divergence, respectively. The $symKL$ is the symmetrized version of the KL divergence. Note that all of the above loss functions are motive by divergences, but they are not divergences since they measure the point-to-point distance while divergences measure the distance between two distributions which lost the data corresponding information.

3.2. The Correlation Coefficients

In statistics, there are two categories of correlation coefficients: linear and non-linear correlation coefficients.

Since the correlation coefficient is used to show how the scores from one measure relate to scores on a second measure from the same group of variables. In our works, the correlation coefficients are used to find basis functions, which contain the same tendency as the evaluation metrics.

3.2.1. Pearson Correlation Coefficient

Pearson Correlation Coefficient (PCC) is a type of linear correlation coefficient. Given a pair of variables (X, Y) the length of each variable is n, Pearson correlation coefficient $\rho_{(X,Y)}$ is calculated by the following formula:

$$\rho^{p}_{(X,Y)} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{13}$$

where $\mathrm{cov}(X,Y)$, and σ_X, σ_Y are the covariance and standard deviation of X and Y.

3.2.2. Spearman Correlation Coefficient

Spearman correlation coefficient (SCC) is a type of non-linear correlation coefficient. SCC is defined as the Pearson correlation coefficient between the rank variables. For X and Y, we first sort them get the rank variables $R(X)$ and $R(Y)$. Then the Spearman correlation coefficient $\rho^{s}_{(X,Y)}$ is calculated as:

$$\rho^{s}_{(X,Y)} \triangleq \rho^{p}_{(R(X),R(Y))} = \frac{\mathrm{cov}(R(X),R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \tag{14}$$

3.2.3. Kendall Correlation Coefficient

Kendall correlation coefficient (KCC) is also a type of non-linear correlation coefficients. For X and Y, we first sort Y by X in ascending order, then count the number of concordant and discordant pairs in the sorted Y, where pair (i, j) is a concordant pair if $i < j$ else it is a discordant pair.

$$\rho^{k}_{(X,Y)} = \frac{C - D}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \tag{15}$$

where C and D are the number of concordant and discordant pairs, $\begin{pmatrix} n \\ 2 \end{pmatrix} = \frac{n(n-1)}{2}$ is the binomial coefficient for the number of ways to choose two items from n items.

All above three correlation coefficients are ranged from -1 to +1 depending on the definition. The physical meaning of the correlation coefficient is that a negative number means the negative correlation between X and Y, and a positive number means the positive correlation, $i.e.$ evaluation metrics tends to increases as the value of loss function is decreased.

The closer the number of the correlation coefficient to the boundary (-1 or 1) the stronger correlation(negative or positive) between X and Y. If the value of correlation coefficient is 0, which means that X and Y have no linear correlation for PCC or non-linear correlation for SCC and KCC.

3.3. The Selection Strategy

For correlation analysis, we build a selection dataset generated by all types of noise in the training set, and SNR in the range [-5, 5] to conduct. The values of proposed basis loss functions and the values of evaluation metrics are calculated on the proposed selection dataset. Then, the corresponding correlation coefficient between basis loss functions and evaluation metrics are calculated by the methods introduced in 3.2. Through the correlation coefficient, we could select the best loss function for training. We analyze all of the listed basis loss functions of section 3.1. Especially, we also combine the $rGKL$ with MSE and JS to produce new substitutions, since their outstanding PCC, SCC, and KCC.

4. EXPERIMENTS AND RESULTS

4.1. Experiments Setup and Datasets

The dataset in our experiments is derived from the TIMIT [22] dataset. The NoiseX-92 [23] dataset is selected as the noise dataset. To reflect the generalization performance, in terms of noise type selection, we use the first half of each noise as the training set noise and the second half as the test set noise. In terms of signal-to-noise ratio (SNR), we use -5, 0 dB on the training set, and on the test set, we use -5, 0, 5 dB, 5 dB is the case of SNR mismatch.

We build a simple mask-based baseline model as our testbed, composed of two layers of BLSTM and a linear layer. The BLSTM layer has 384 neurons and the linear layer is used as the output layer to modify the output shape of the BLSTM. To prevent over-fitting, dropout is added to each layer of BLSTM, where the rate is 0.4. Sigmoid is used as the activation function of the output layer.

The sampling rate of each utterance in this experiment is 16 kHz, and the amplitude spectrum is used as the feature. The energy of each utterance is normalized, a 20 ms Hamming window with a 10 ms window shift is used to perform the short-time Fourier transform on the noisy speech, and after obtaining the amplitude spectrum, it is fed into the network. To illustrate the experimental results, all experiments with this article use the same model and parameters for training. Using Adam as the optimizer, the learning rate is 0.001, the models have trained 200 epochs, and the batch size is set to 32.

The performance of the single-channel speech enhancement system is evaluated with four objective metrics: PESQ, STOI, Signal-to-Noise Ratios (SNR) and Signal-to-Distortion Ratio (SDR) [24]. For all indicators, a higher score means better performance.

4.2. Experiments Results

The experimental results are shown in Tab.1, where the best performance in each situation are highlighted in bold. On the left part of Table 1, $\mathbf{PCC}_{sum}$, $\mathbf{SCC}_{sum}$, and $\mathbf{KCC}_{sum}$ means the sum of the corresponding correlation coefficients between the loss function and each evaluation metrics.

Table 1. Speech enhancement performance of different loss functions on test set at -5, 0, 5 dB.

Loss	$\mathbf{PCC}_{sum}$	$\mathbf{SCC}_{sum}$	$\mathbf{KCC}_{sum}$	SNR of Test Set (dB)											
				-5				0				5			
				STOI	PESQ	SNR	SDR	STOI	PESQ	SNR	SDR	STOI	PESQ	SNR	SDR
mix	-	-	-	0.6218	1.5169	-5.0000	-4.8282	0.7351	1.8744	0.0000	0.0934	0.8171	2.1771	5.0000	5.0704
MSE	-2.1547	-2.9824	-2.2250	0.8210	2.3860	6.9691	6.9792	0.8830	2.7755	10.4393	10.9129	0.9170	3.0278	13.3688	14.1970
symKL	-2.9562	-2.9695	-2.2308	0.8214	2.4030	7.0090	7.0881	0.8830	2.7861	10.4284	10.9239	0.9173	3.0690	13.3770	14.2159
GKL	0.3065	0.2705	0.3402	0.8199	2.3567	6.7550	6.6564	0.8816	2.7471	10.2004	10.5997	0.9172	3.0222	13.1907	13.9767
rGKL	-2.9710	**-3.1483**	**-2.3873**	0.8209	2.4155	**7.1340**	**7.3421**	0.8823	2.8027	10.5006	11.0238	0.9059	3.0933	12.7880	14.0382
JS	-2.9391	-2.9418	-2.2042	0.8217	2.3835	6.8955	6.8706	0.8826	2.8705	10.3015	10.7118	0.9173	3.0645	13.3122	14.0969
IS	0.3784	0.8135	0.5521	0.7833	2.0816	5.6359	5.0566	0.8529	2.4632	8.9272	9.0457	0.8945	2.7329	11.9814	12.5536
rIS	-1.6508	-2.6422	-1.9300	0.7876	2.2778	6.2888	7.0316	0.8556	2.6856	9.6037	10.5736	0.8944	2.9788	12.4423	13.6290
rGKL+MSE	-2.4865	-3.1387	-2.3768	0.8218	2.3815	7.0348	7.0909	0.8832	2.7980	10.4662	10.9443	0.9171	3.0738	13.3666	14.1908
rGKL+JS	**-2.9966**	-3.1148	-2.3644	**0.8242**	**2.4222**	7.1054	7.32736	**0.8840**	**2.8097**	**10.5299**	**11.0665**	**0.9188**	**3.1053**	**13.4533**	**14.2904**

Follow the $\mathbf{PCC}_{sum}$, the combination of loss function **rGKL + JS** has the highest linear correlation coefficient with evaluation metrics. Meanwhile, the loss function **rGKL+JS** achieves the best performance on all metrics at the SNR in 0 and 5 dB and has the best performance at metrics of STOI and PESQ at the test set of -5dB. It demonstrates that the loss function combination **rGKL+JS** is more suitable for speech enhancement system than the other basis loss functions. However, the **rGKL** loss function which has the biggest $\mathbf{SCC}_{sum}$ and $\mathbf{KCC}_{sum}$ underperforms the **rGKL+JS**, which indicates that the linear correlation coefficient may be more effective than the non-linear ones in loss selection.

Table 2. Pearson Correlation Coefficient (PCC)

Loss	SUM	STOI	PESQ	SNR	SDR
MSE	-2.1547	-0.4823	-0.5587	-0.5894	-0.5243
symKL	-2.9562	-0.7317	-0.7834	-0.7249	-0.7162
GKL	0.3065	-0.0702	0.0239	0.2156	0.1372
rGKL	-2.9710	-0.7562	-0.8013	-0.7037	-0.7098
JS	-2.9391	-0.7127	-0.7730	**-0.7336**	**-0.7198**
IS	0.3784	0.1316	0.1131	0.0491	0.0846
rIS	-1.6508	-0.4851	-0.4932	-0.3239	-0.3486
rGKL+MSE	-2.4865	-0.5785	-0.6552	-0.6539	-0.6019
rGKL+JS	**-2.9966**	**-0.7565**	**-0.8048**	-0.7163	-0.7190

4.3. Further Analysis in Correlation Coefficient

To further understand the behaviors of PCC, SCC, and KCC, we show the three different correlation coefficients between loss functions and evaluation metrics in Table 2, 3 and 4. In Table 2, **rGKL+JS** gets the highest PCC in STOI and PESQ, but in SNR and SDR it performs slightly weak. This may be the reason why the results of loss function **rGKL+JS** is not optimal in these evaluation metrics scores at low test set SNR in Table 1. Following the Tab.3 are Tab.4, the **rGKL** could be the best loss function in speech enhancement system. However, sometimes the **rGKL** can no even beat **MSE**, although it outperforms **rGKL+JS** in SNR and SDR in -5dB. This results further consistent with our previous conclusions that the linear correlation coefficient higher the better performance of loss function.

5. CONCLUSIONS

This paper studies the loss-metric mismatch problem in the supervised single-channel speech enhancement system. We propose a simple yet efficient strategy to generate suitable loss functions for single-channel speech enhancement system by approximating the non-differentiable evaluation metric by a set of basis functions and their linear combination. By alleviating the loss-metric mismatch problem, the experimental results demonstrate that the loss function generated by our method helps the speech enhancement system achieve remarkable performance in most evaluation metrics than the traditional empirically selected ones.

Table 3. Spearman Correlation Coefficient (SCC)

Loss	SUM	STOI	PESQ	SNR	SDR
MSE	-2.9824	-0.6618	-0.7310	-0.8153	-0.7743
symKL	-2.9695	-0.7361	-0.7766	-0.7385	-0.7183
GKL	0.3402	0.0250	0.0091	0.1982	0.1079
rGKL	**-3.1483**	**-0.7734**	**-0.8361**	-0.7664	-0.7624
JS	-2.9418	-0.7238	-0.7597	-0.7413	-0.7170
IS	0.8135	0.1617	0.2249	0.1825	0.2444
rIS	-2.6422	-0.7027	-0.7628	-0.5725	-0.6042
rGKL+MSE	-3.1387	-0.7169	-0.7924	**-0.8307**	**-0.7987**
rGKL+JS	-3.1148	-0.7678	-0.8257	-0.7644	-0.7569

Table 4. Kendall Correlation Coefficient (KCC)

Loss	SUM	STOI	PESQ	SNR	SDR
MSE	-2.2250	-0.4775	-0.5407	-0.6260	-0.5807
symKL	-2.2308	-0.5516	-0.5912	-0.5530	-0.5350
GKL	0.2705	0.0207	0.0148	0.1480	0.0870
rGKL	**-2.3873**	**-0.5847**	**-0.6526**	-0.5768	-0.5732
JS	-2.2042	-0.5408	-0.5747	-0.5555	-0.5332
IS	0.5521	0.1125	0.1512	0.1232	0.1652
rIS	-1.9300	-0.5210	-0.5721	-0.4055	-0.4314
rGKL+MSE	-2.3768	-0.5267	-0.6008	**-0.6424**	**-0.6069**
rGKL+JS	-2.3644	-0.5798	-0.6406	-0.5755	-0.5685

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61876214, 61866030, 62066033), National Key Research and Development Program of China (Grant No. 2018YFE0122900) and Applied Technology Research and Development Program of Inner Mongolia Autonomous Region (Grant No. 2019GG372, 2020GG0046, 2021GG0158, 2020PT0002)

7. REFERENCES

[1] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[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] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," vol. 2, pp. 629–632 vol. 2, 1996.

[4] N. Mohammadiha, P. Smaragdis, et al., "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[5] DeLiang Wang and Guy J. Brown, "Computational auditory scene analysis: Principles, algorithms, and applications," *Wiley-IEEE Press*, 2006.

[6] DeLiang Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, pp. 181–197. 2005.

[7] Chen Huang, Shuangfei Zhai, et al., "Addressing the loss-metric mismatch with adaptive loss alignment," in *ICML*, 2019, vol. 97, pp. 2891–2900.

[8] C. H. Taal, R. C. Hendriks, et al., "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[9] A. W. Rix, J. G. Beerends, et al., "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," vol. 2, pp. 749–752 vol.2, 2001.

[10] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 47–56, 2011.

[11] P.C. Loizou, "Speech enhancement: theory and practice: Crc press," 2013.

[12] Y. Shen, C. Huang, et al., "Reinforcement learning based speech enhancement for robust speech recognition," in *ICASSP*, 2019, pp. 6750–6754.

[13] L. P. Kaelbling, M. L. Littman, et al., "Reinforcement Learning: A Survey," *arXiv e-prints*, p. cs/9605103, apr 1996.

[14] S. Fu, C. Liao, et al., "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality," *IEEE Signal Processing Letters*, vol. 27, pp. 26–30, 2020.

[15] H. Zhang, X. Zhang, et al., "Training supervised speech separation system to improve stoi and pesq directly," in *ICASSP*, 2018, pp. 5374–5378.

[16] Morten Kolbæk, Zheng-Hua Tan, et al., "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *ICASSP*, 2018, pp. 5059–5063.

[17] Joseph Kreimer and Reuven Y. Rubinstein, "Nondifferentiable optimization via smooth approximation: General analytical approach," *Ann. Oper. Res.*, vol. 39, no. 1, pp. 97–119, 1992.

[18] Yun Liu, Hui Zhang, et al., "Investigation of cost function for supervised monaural speech separation," in *Interspeech*, 2019, pp. 3178–3182.

[19] Y. Xu, J. Du, et al., "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2015.

[20] Y. Xu, J. Du, et al., "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

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

[22] J. W. Lyons, "Darpa timit acoustic-phonetic continuous speech corpus," *National Institute of Standards and Technology*, 1993.

[23] Andrew Varga and Herman J.M. 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.

[24] Emmanuel Vincent, Rémi Gribonval, et al., "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.