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

Investigation of Cost Function for Supervised Monaural Speech Separation

Yun Liu[1], *Hui Zhang*[1], *Xueliang Zhang*[1], *Yuhang Cao*[2]

[1]Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Inner Mongolia University, Hohhot, China
[2]Beijing Unisound Information Technology Co. Ltd., China

liuyun.nogizaka@qq.com, alzhu.san@163.com, cszxl@imu.edu.cn, caoyuhang@unisound.com

Abstract

Speech separation aims to improve the speech quality of noisy speech. Deep learning based speech separation methods usually use mean square error (MSE) as the cost function, which measures the distance between model output and training target. However, the MSE does not match the evaluation metrics perfectly. Optimizing the MSE does not directly lead to improvement in the commonly used metrics, such as short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ), signal-to-noise ratio (SNR) and source-to-distortion ratio (SDR). In this study, we inspect some other cost function candidates which are based on divergence, e.g., Kullback-Leibler and Itakura-Saito divergence. A conjecture about the correlation between cost function and evaluation metrics is proposed and examined to explain why these cost functions behave differently. On the basis of the proposed conjecture, the optimal cost function candidate is selected. The experimental results validate our conjecture.

Index Terms: divergence, deep neural networks, cost function, speech separation

1. Introduction

The purpose of speech separation is to recover a target signal from a mixture of background noise. Speech separation has a wide range of prospective applications, such as robust automatic speech recognition (ASR), mobile speech communication and speaker recognition [1].

Traditionally, the speech separation problem has been addressed using many methods, such as spectral subtraction [2], Wiener filtering [3], statistical model based methods [4] and nonnegative matrix factorization [5]. However, these conventional methods assume that noise is stationary, but most of the noises in the real-world environment are nonstationary, which leads to difficulty in using these methods in real-world application scenarios. Subsequently, with the development of computational auditory scene analysis (CASA) [6], the speech separation problem has been formulated as a supervised learning problem that leads to an effective solution to nonstationary noise condition.

Many studies related to the field of supervised speech separation have been conducted. Researchers devoted many works to features [7], training targets [8, 9] and models. To obtain the desired result under a complicated noise environment, many superior models have been employed and have obtained considerable performance improvements. Deep neural networks (DNNs), long short term memory networks (LSTMs) and generative adversarial networks (GANs) have been widely used and have exhibited powerful modeling capabilities in speech separation [9–11]. However, only a few works focus on the cost function.

In the field of supervised speech separation, a learning machine is employed to learn a mapping function from noisy acoustic features to time-frequency (T-F) maskings or spectral representations of target speech, where these two main methods can be generally referred to the masking-based and mapping-based methods. The masking-based method minimizes the mean square error (MSE) between the estimation and the ideal mask target, and the mapping-based method minimizes the MSE between the estimation and the target clean spectrum. Signal approximation (SA) methods combine the masking-based and mapping-based methods [10, 12]. In the SA methods, a learning machine is trained to estimate the T-F mask, but their cost function is the MSE between the estimation and the target clean spectrum.

In these studies, the MSE is the most commonly used cost function, although it is not perfect. Many drawbacks have been reported. For instance, the MSE does not consider the difference between high-frequency and low-frequency energies at each T-F unit. The MSE only measures the local relationship and it does not reflect the global condition. Recently, several studies proposed and focused on the cost function of supervised speech separation , which shows considerable improvements compared with the MSE. Zhang [13] used a gradient approximation method to calculate the gradient of the short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) loss function, which trains a supervised speech separation system to improve the STOI and PESQ directly. The simplified STOI was also be used as the cost function in [14, 15] and determined to be better than conventional MSE-optimized speech separation methods because the consistency between the training and the evaluation targets. Chai [16] used a weighted MSE loss function in the maximum likelihood (ML) approach which assumes that the prediction error vector of DNN follows a Gaussian distribution. Other studies [17, 18] also investigated the use of source to distortion ratio (SDR) as the cost function.

Cost function, such as MSE, is used to measure the distance between model output and training target. The MSE is not the only option. We were inspired by the work presented in [19] which proposed a nonnegative matrix factorization (NMF) algorithm with Itakura-Saito (IS) divergence. The IS divergence can measure the goodness of fit between two spectra and is considered to be better than MSE. The IS divergence is scale-invariant, that is, low energy components bear the same relative importance as high energy components. Therefore, in addition to the MSE, the IS divergence can be another cost function candidate.

The IS divergence is also not the only option. The IS divergence can be categorized under the class of Bregman divergences and is a limit case of the β-divergence. Other divergences, such as Kullback-Leibler (KL) divergence, also

http://dx.doi.org/10.21437/Interspeech.2019-1897

belong to β-divergence, which has been widely used in image processing. Therefore, we will consider a broad range of divergence-based cost functions.

Of the numerous cost function candidates, which is the best? In this study, we try to answer this question.

For supervised speech separation, the key problem of the cost function is the training-evaluation mismatch problem, in which the cost function does not match the evaluation metrics. For example, minimizing the MSE does not have a direct connection with the improvement in STOI, PESQ, signal-to-noise ratio (SNR) and SDR. The efforts to optimize the cost function may not improve the evaluation metrics. Therefore, we propose that the degree of correlation between cost function and evaluation metrics is an indicator to the cost function's quality. A high degree of correlation may lead to a good cost function. Data analysis and experiments are conducted to examine the proposed conjecture. The optimal cost function selected by the proposed method exhibits an improvement compared with the most commonly used MSE.

2. Divergence-based Speech Separation

2.1. Supervised speech separations

Given a time-domain speech signal s and noise n, we can obtain noisy speech $c = s + n$. We decompose them into the T-F domain via short time Fourier transform and obtain the complex spectra S, N and $C = S + N$. Acoustic features are extracted from the noisy speech signal as input to a neural network. The T-F mask or target magnitude spectrum is used as output. After training, we can obtain estimated mask or magnitude spectrum only from noisy speech. Then, the estimated speech signal can be recovered from the estimated magnitude spectrum together with the noisy phase spectrum.

The MSE is commonly used as the cost function in this framework and measures the error between the network output and the training target. The MSE is a distance in Euclidean space, which can be expressed as follows:

$$MSE(x, y) = |x - y|^2 \tag{1}$$

where x and y denote the true and estimated output, respectively. The mapping-based method trains a network to minimize the MSE between the target magnitude spectrum $|S|$ and estimated spectrum $|\hat{S}|$, where the hat mark means "estimated". Meanwhile, the masking-based method trains a network to minimize the MSE between target and estimated mask. Spectral magnitude mask (SMM) [8] is usually used as training target:

$$SMM = \frac{|S|}{|C|} \tag{2}$$

where $|S|$ and $|C|$ denote the speech magnitude spectrum and noisy magnitude spectrum, respectively.

The SA methods combine the masking-based and mapping-based methods, where a learning machine is trained to estimate these T-F masks. However, their cost function is the MSE of the estimated magnitude spectrum and target magnitude spectrum. For magnitude spectrum approximation (MSA) [12], we define the cost function as follows:

$$E_{SA} = |\hat{M} * |C| - |S||^2 \tag{3}$$

where $\hat{M}$ is the estimated mask. The estimated speech $|\hat{S}|$ is reconstructed by $|\hat{S}| = \hat{M} * |C|$, where $*$ denotes the element-wise multiplication. In the MSA methods, a network is trained to minimize the MSE between $\hat{M} * |C|$ and $|S|$.

2.2. Divergence-based cost function and evaluation metrics

In machine learning, we often need to use the estimated distribution Q to approximate the target distribution P. We hope to derive the cost function $D(P, Q)$, which is usually called divergence, to compute the distance from Q to P. With the decrease in the cost function, the estimated distribution Q will become more similar to the target distribution P.

Many types of divergence are defined. We list some of them, as follows:

$$KL\,(x, y) = x \cdot \log \frac{x}{y} \tag{4}$$

$$symKL\,(x, y) = x \cdot \log \frac{x}{y} + y \cdot \log \frac{y}{x} \tag{5}$$

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

$$rGKL\,(x, y) = y \cdot \log \frac{y}{x} - (y - x) \tag{7}$$

$$JS\,(x, y) = \frac{1}{2}\left(x \cdot \log \frac{2x}{x + y} + y \cdot \log \frac{2y}{x + y}\right) \tag{8}$$

$$IS\,(x, y) = \frac{x}{y} - \log \frac{x}{y} - 1 \tag{9}$$

where KL is the KL divergence; GKL is the generalized form of the KL divergence and $rGKL$ is the reverse version of the GKL. $symKL$ is the symmetrized version of the KL divergence or rGKL divergence. JS is the Jensen-Shannon divergence, which is a special symmetrized and smoothed version of the KL divergence. IS is the IS divergence, which was obtained from the maximum likelihood (ML) estimation of short-time speech spectra under autoregressive modeling. All of these divergences can be used as the cost function in supervised speech separation. In the masking-based method, we set x as the target mask and y as the estimated mask. In the mapping-based method, we set $x = |S|$ and $y = |\hat{S}|$. In the MSA method, we set $x = |S|$ and $y = \hat{M} * |C|$.

From the analysis of all of the cost function candidates, we determined that all of them can be rewritten in the form of $w \cdot b$, where w is the weight vector, and b is the vector consisting of a batch of basis functions.

$$b = \begin{bmatrix} x - y \\ |x - y|^2 \\ \frac{x}{y} \\ \frac{y}{x} \\ \log(\frac{x}{y}) \\ \log(\frac{y}{x}) \\ x\log(\frac{x}{y}) \\ y\log(\frac{y}{x}) \\ x\log(\frac{2x}{x+y}) \\ y\log(\frac{2y}{x+y}) \\ 1 \end{bmatrix} \tag{10}$$

Then all cost function candidates can be rewritten as follows:

$$MSE\,(x, y) = [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0] \cdot b \tag{11}$$

$$KL\,(x, y) = [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0] \cdot b \tag{12}$$

$$symKL\,(x, y) = [0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0] \cdot b \tag{13}$$

$$GKL\,(x, y) = [-1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0] \cdot b \tag{14}$$

$$rGKL\,(x, y) = [1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0] \cdot b \tag{15}$$

$$JS\,(x, y) = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0.5, 0.5, 0] \cdot b \tag{16}$$

$$IS\,(x, y) = [0, 0, 1, 0, -1, 0, 0, 0, 0, 0, -1] \cdot b \tag{17}$$

Using this format, we can define other divergence-like cost functions, such as:

$$rIS\,(x,y) = [0,1,0,-1,0,0,0,0,0,0,1]\cdot b \quad (18)$$

$$rGKL+MSE\,(x,y) = [-1,1,0,0,0,0,0,1,0,0,0,0]\cdot b \quad (19)$$

$$rGKL+JS\,(x,y) = [-1,0,0,0,0,0,0,1,0.5,0.5,0,0]\cdot b \quad (20)$$

where rIS is the reversed IS divergence. $rGKL+MSE$ and $rGKL+JS$ are the combinations of $rGKL$ with MSE or JS, respectively.

The speech separation performance is usually evaluated in terms of STOI [20], PESQ [21], SNR, and SDR. For all metrics, a higher score means a better performance. These evaluation metrics are commonly used to evaluate the speech intelligibility and quality. In [22], it is indicated that the SDR has an obvious correlation with the word error rate of a speech recognition system.

2.3. Evaluation metrics vs. cost function

Supervised speech separation trains a learning machine to improve the evaluation metrics by optimizing the cost function. A clean mismatch exists between the evaluation metrics and the cost function. The efficiency E of the cost function can be defined as follows:

$$E = \frac{\Delta_{metric}}{\Delta_{cost}} \quad (21)$$

where Δ_{metric} is the change of the evaluation metric, and Δ_{cost} is the change of the cost function. The efficiency E measures the sensitivity to the change of the evaluation metrics with respect to the change of the cost function. Ideally, any change of the cost function leads to a corresponding change of the evaluation metric. Therefore, if the cost function and evaluation metrics have a strong linear correlation, than the cost function will have a high efficiency. The degree of linear correlation can be evaluated with the following correlation coefficient:

$$\rho(x,y) = \frac{S(x,y)}{S(x)\cdot S(y)} \quad (22)$$

where $S(x)$ and $S(y)$ are the sample standard deviations, and $S(x,y)$ is the sample covariance. The correlation coefficient value ranges between -1 and $+1$. A value of -1 indicates a strong negative correlation and a value of $+1$ indicates a strong positive correlation. In the context of optimization, we want a low cost function value that correspond to a high evaluation metric. Thus, we want the cost function to have a near -1 correlation coefficient with the evaluation metric. To increase the correlation between the cost function and all of the evaluation metrics, we define the quality function $q(w)$ for any cost function defined in the form of $w\cdot b$:

$$\begin{aligned} q(w) = {} & \rho(w\cdot b, STOI) + \rho(w\cdot b, PESQ) \\ & +\rho(w\cdot b, SNR) + \rho(w\cdot b, SDR) \end{aligned} \quad (23)$$

A low $q(w)$ value indicates high quality because $\rho(x,y)$ should be negative.

In this manner, we can also identify the best configuration of w to minimize $q(w)$, which is the best cost function in this problem setting.

$$\arg\min_{w} q\,(w) \quad (24)$$

We will examine the proposed conjecture through the data analysis and experiments presented in the following section.

3. Experiments and Results

3.1. Dataset and system setup

The dataset evaluated in our experiment is derived from the TIMIT dataset [23]. The dataset generation method is the same as used in [8]. In our experiment, 2000 utterances from the TIMIT training set are randomly selected as the target speech for training. All 192 utterances from the TIMIT core test set are selected for the test. Five types of noises are used for training, that is a speech-shaped noise (SSN) and four other types of noises selected from the NOISEX database [24], which include babble noise, factory noise, destroy engine noise, and destroyer operations room noise. Four other types of new noises from the CHiME-4 dataset, which consists of bus noise, cafe noise, street noise and pedestrian noise, are used for noise-unmatched condition test. The training set is built by mixing all of the target speech and noises at -5 and 0 dB SNR. The test set is built by mixing all of the target speech and noises at -5, 0, and 5 dB SNR, where 5 dB SNR is the SNR-unmatched condition. The input features are a complementary set that includes four features are extracted from a 64-channel Gammatone filterbank, that is, amplitude modulation spectrogram (AMS), mel-frequency cepstral coefficient (MFCC), relative spectral transforms perceptual linear prediction (RASTA-PLP) and cochleagram response, as well as their deltas [9]. Mean and variance normalization is applied to these features before feeding them into the bidirectional long short-term memory (BLSTM) network. We use a two-layer BLSTM with 384 cells and a fully connected layer as the output layer. Dropout regularization is applied on each BLSTM layer to prevent overfitting and the dropout rate is 0.4. All of the experiments use the same BLSTM structure as the model. This network was trained using different cost functions.

All used spectra are generated by resampling signals into 16 kHz, and dividing into frames using a 20 ms Hamming window with 10 ms overlap. Adam is used for optimization and the initial leaning rate is 0.001.

To evaluate performance, we built a speech separation system in the manner of MSA. The training target and model output are clipped to $(1\times10^{-6}, 10)$ when calculating the loss, where 1×10^{-6} is a small number to avoid dividing by 0, and 10 is a arbitrarily selected number to avoid a large dynamic range.

3.2. Quality of the cost function

We use the training set as our data source, we calculate all of the evaluation metrics and basis functions in b. Then, the results of b are used to form the cost function results. Table. 2 lists the correlation coefficient of the cost function with respect to STOI, PESQ, SNR and SDR. By summing up each line, the quality of the cost functions is evaluated using the sum of correlation coefficients.

Following the proposed conjecture, the JS, symKL, rGKL and rGKL+JS will be the first-class cost function, others will perform worse, and IS with its reversed version will be the worst choice.

By solving the problem defined in (24) using the gradient descent method of numerical derivation, a optimal solution is obtained as follows:

$$\begin{aligned} w = [& 0.00001519925, 0.00065652066, 0.00002196230, \\ & -0.00019905547, 0.00069584670, -0.00069584670, \\ & 0.00006584202, -0.00001574538, 0.50000774714, \\ & 0.49999194313, 0.00000000000] \end{aligned}$$

Table 1: *Speech separation performance of different cost function at -5,0,5 dB.*

cost function	SNR(dB)											
	-5				0				5			
	STOI(%)	**PESQ**	**SNR**	**SDR**	**STOI(%)**	**PESQ**	**SNR**	**SDR**	**STOI(%)**	**PESQ**	**SNR**	**SDR**
mix	64.1	1.58	-5.00	-4.78	73.4	1.94	0.00	0.11	81.8	2.29	5.00	5.08
MSE	78.1	2.26	6.01	6.21	85.4	2.61	9.22	9.86	90.1	2.93	12.34	13.33
symKL	77.9	2.25	6.24	6.35	85.6	2.65	9.32	9.92	90.2	2.99	12.23	13.21
GKL	76.7	2.20	5.41	5.69	84.2	2.56	8.34	9.26	89.0	2.88	11.11	12.73
rGKL	78.3	2.28	6.38	6.45	85.6	2.66	9.27	9.85	90.2	3.00	12.24	13.37
JS	**79.1**	**2.29**	6.13	6.35	**86.3**	**2.68**	9.11	9.81	**90.6**	**3.01**	12.01	13.15
IS	75.2	2.04	3.47	3.85	82.7	2.41	6.28	7.67	87.9	2.76	9.04	11.50
rIS	74.1	2.05	5.61	5.47	82.1	2.49	8.35	9.05	87.0	2.84	10.71	12.21
rGKL+MSE	78.4	2.27	6.41	6.40	85.8	2.67	**9.42**	9.90	90.2	3.00	**12.55**	13.39
JS+rGKL	78.7	2.27	**6.45**	**6.67**	86.0	2.67	9.41	**10.06**	90.5	**3.01**	12.41	**13.42**

Table 2: *Correlation coefficient of cost function with STOI, PESQ, SNR and SDR.*

Cost func.	STOI(%)	PESQ	SNR	SDR	$q(w)$
MSE	-0.6478	-0.6678	-0.8161	-0.8146	-2.9463
symKL	-0.7920	-0.7892	-0.8568	-0.8550	-3.2931
GKL	-0.7852	-0.7841	-0.8572	-0.8555	-3.2819
rGKL	-0.8097	-0.8005	-0.8471	-0.8452	**-3.3025**
JS	-0.8082	-0.7995	-0.8488	-0.8469	**-3.3010**
IS	-0.3127	-0.2781	-0.3212	-0.3211	-1.2331
rIS	-0.5647	-0.5986	-0.7500	-0.7488	-2.6621
rGKL+MSE	-0.7235	-0.7355	-0.8574	-0.8557	-3.1721
rGKL+JS	-0.7897	-0.7875	-0.8570	-0.8553	-3.2894

Table 3: *Speech separation relative average increase performance of different cost function for MSA.*

Cost function	STOI(%)	PESQ	SNR	SDR
MSE	11.43	0.66	9.19	9.66
symKL	11.47	0.69	9.26	9.69
GKL	10.20	0.61	8.29	9.09
rGKL	11.60	0.71	9.30	9.75
JS	**12.23**	**0.72**	9.08	9.63
IS	8.83	0.47	6.26	7.54
rIS	7.97	0.52	8.22	8.77
rGKL+MSE	11.70	0.71	**9.46**	9.76
rGKL+JS	11.97	**0.72**	9.42	**9.91**

The solution is very similar to the JS. Meanwhile, the quality function $q(w)$ is -3.3070, which is sightly higher than the value of JS (-3.3010). Considering the sampling variance and simplicity of the definition, we can regard the JS divergence as the optimal training cost function under the current setting.

3.3. Evaluation with different cost functions

The experimental results are listed in Table 1, and Table 3. Table 1 shows the performance under different SNR conditions. Table 3 summarizes all of the SNR conditions and shows the improvement relative to the unprocessed signal.

From Table 1 and Table 3, we can observe that the results of JS+rGKL generally exhibit the best performance. In terms of PESQ, the cost function which has the JS or rGKL, shows a obvious advantage over the MSE. The combined cost function which has the reverse GKL with MSE, exhibits an increase in SNR, indicating that the MSE has a regulating effect on SNR performance. The GKL shows worse results than the reverse GKL and IS shows worse results than reverse IS, which indicated that a reverse version of divergence may be more suitable for a fitness between two spectra. The JS divergence performs better than the MSE, except for the SNR and SDR. Notably, it obtains the highest STOI and PESQ gains compared with other cost functions in general. However, the poor performance of IS or rIS may have resulted from the large dynamic range of x/y or y/x, which leads to instability when training.

Comparing Table 3 with Table 2, we can see a clear correspondence between the correlation coefficients and the relative improvement. The high correlation score leads to a high speech separation improvement in general as the proposed conjecture suggested.

4. Conclusions

In this study we proposed a method that can select the optimal cost function from candidates based on the correlation between cost function and evaluation metrics. The higher correlation leads to a better performance in general. In the future, more cost functions and more evaluation metrics will be added to the candidates. For example, some powerful distances, such as Wasserstein distance that has an advantage over the KL and JS divergence, can also be employed as cost function. On the basis of the proposed method, better cost functions over the commonly used MSE will be detected in the supervised separation system, in the robust ASR system and in other systems facing the training-evaluation mismatch problem.

5. Acknowledgements

This research was supported in part by the China national nature science foundation (No. 61876214, No. 61866030).

6. References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, Oct 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 *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2. IEEE, 1996, pp. 629–632.

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

[5] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[6] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[7] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.

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

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

[10] 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.

[11] A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5414–5418.

[12] 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.

[13] H. Zhang, X. Zhang, and G. Gao, "Training supervised speech separation system to improve stoi and pesq directly," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5374–5378.

[14] M. Kolbcek, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5059–5063.

[15] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1570–1584, 2018.

[16] 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.

[17] S. Venkataramani, R. Higa, and P. Smaragdis, "Performance based cost functions for end-to-end speech separation," *arXiv preprint arXiv:1806.00511*, 2018.

[18] H. Nakajima, Y. Takahashi, K. Kondo, and Y. Hisaminato, "Monaural source enhancement maximizing source-to-distortion ratio via automatic differentiation," *arXiv preprint arXiv:1806.05791*, 2018.

[19] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

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

[21] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech and Signal Processing (ICASSP), 2001 IEEE International Conference on*, 2001, pp. 749–752.

[22] 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.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1*, 1993, vol. 93.

[24] 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.