# Circuits, Systems, and Signal Processing


## Correntropy based Multi-objective Multi-channel Speech Enhancement

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | CSSP-D-21-00332R1 |
| **Full Title:** | Correntropy based Multi-objective Multi-channel Speech Enhancement |
| **Article Type:** | Original Research |
| **Keywords:** | Speech enhancement, microphone array, correntropy, neural networks, masking |
| **Abstract:** | Although deep learning-based methods have greatly advanced the speech enhancement, their performance are intensively degraded under the non-Gaussian noises. To combat the problem, a correntropy based multi-objective multi-channel speech enhancement method is proposed. First, the log-power spectra (LPS) of multichannel noisy speech are feed to the bidirectional long short-term memory (BiLSTM) network with the aim of predicting the intermediate ideal ratio mask (IRM) and LPS of clean speech in each channel. Then, the intermediate LPS and IRM features obtained from each channel are integrated into a single-channel LPS and IRM feature by using a fusion layer, respectively. Next, the two single-channel features are fused into a single-channel LPS and feed to the deep neural network to learn the relationship between the predicted LPS and the clean speech LPS. Finally, a loss function for network training is constructed by correntropy to improve the performance of overall network. Experimental results show that the proposed method achieves significant improvements in suppressing non-Gaussian noises and reverberations, and has good robustness to different noises, signal-noise-ratios (SNRs) and source-array distances. |
| **Response to Reviewers:** | Correntropy based Multi-objective Multi-channel Speech Enhancement<br>(ID: CSSP-D-21-00332)<br>Xingyue Cui, Zhe Chen, Fuliang Yin, Xianfa Xu<br><br>The Point-by-point Responses to the Reviewer's Comments<br>We would like to express our appreciation to the reviewers for providing us with valuable comments for improving this manuscript. In the following, we present the point-by-point replies to the reviewers' comments.<br><br>Reviewer1:<br><br>Comments No.1: However, the authors show the experimental results, which evident the claim, but an analytical claim for the same is missing. Better to provide a subsection, and explain why correntropy loss performs better with non Gaussian noise (appropriate reference for the same is also missing). The computation of IRM features is also not so clear, expected to provide a block diagram for the same.<br><br>Answer:<br>Thanks for the reviewer's valuable advice. Generally, non-Gaussian noises, especially impulse noise, are prone to large outliers during training process. Correntropy loss has the capability to suppress the outliers because of the nonlinear kernel function, such as Gaussian kernel function. According to the characteristic of Gaussian function, the lager the outliers, the better the suppression. In contrast, MSE loss would amplify the outliers during training, and is not optimal in the case of asymmetric error distribution, non-zero center and outliers. In other words, nonlinear kernel function is the key factor that makes correntropy robust to outliers, and it is the reason why correntropy loss performs better with non-Gaussian noise.<br>According to the reviewer's comment, we have added the above reason and reference in the revised manuscript. Please refer to the fourth paragraph of Subsection 2.5. Besides, a block diagram of the computation of IRM feature has been also added in the revised manuscript. Please refer to Fig. 1 of Subsection 2.3.<br><br>Comments No.2: The motivation for the work is well written, If the authors can summarize the motivation through a table or Block diagram, it makes the article more readable. |

Answer:
According to the reviewer’s advice, we have summarized the motivation through a table. Please refer to the Table 1 of Section 1 in the revised manuscript.

Comments No.3: "First, the log-power spectra (LPS) of multichannel noisy speech are used as the input of the bi-directional long short-term memory (BiLSTM) network to predict the ideal ratio mask (IRM) and LPS of clean speech in each channel.", In the abstract, the author may re-frame the sentence, it is not so clear.

Answer:
According to the reviewer’s advice, we have modified the sentence “First, the log-power spectra (LPS) of multichannel noisy speech are used as the input of the bi-directional long short-term memory (BiLSTM) network to predict the ideal ratio mask (IRM) and LPS of clean speech in each channel.” to “First, the log-power spectra (LPS) of multichannel noisy speech are feed to the bidirectional long short-term memory (BiLSTM) network with the aim of predicting the intermediate ideal ratio mask (IRM) and LPS of clean speech in each channel.” with red font in the revised manuscript.

Comments No.4: Page no. 4, line 26, the abbreviation of SFTF is missing.

Answer:
Sorry for our carelessness. We have added the abbreviation of SFTF with red font in the revised manuscript.

Comments No.5: In section 2.3, it will be better to explain why clean speech signal magnitude is less than the magnitude of noisy signal. (Ideally)

Answer:
Thanks for the reviewer’s comment. Section 2.3 describes ideal ratio mask which is defined as  in Eq.(4), where S(n,k) and X(n,k) denote the clean- and noisy- spectral magnitudes within a T-F unit, respectively. In fact, S(n,k) is not always less than X(n,k), and the values of IRM may be greater than one. However, a larger value may lead to instability of backpropagation during training, thus, the maximum value of IRM is set to 1 in our method, i.e. S(n,k) has to be no greater than X(n,k). We have explained the reason in the revised manuscript, please refer to the sentences below Eq.(4) in Section 2.3.

Reviewer2:
Comments No.1: The main contribution of the manuscript is exploring Correntropy as the loss function for non-Gaussian distribution. The selection of the kernel size is very crucial to get improve performance compared to other methods. In the manuscript the authors have chosen kernel size based on PESQ and different SNR. However, I have a concern with the selection of the kernel function i.e., Gaussian kernel in the manuscript. Although Gaussian kernel is desirable, but the correntropy induced metric varies with the data and kernel size. In the literature, methods are presented to avoid choosing kernel size and match the sample distribution. Therefore, a proper justification for this concern is expected from the authors.

Answer:
The reviewer raises an interesting concern. Reference [A1] ([30] in revised manuscript) first proposed the concept of correntropy which is based on the kernel method and information theoretic learning theory. It is pointed that the basic idea of kernel method is to transform the data xi from the input space to a high dimensional feature space of vectors  , where the inner products can be computed using a positive definite kernel function satisfying Mercer’s conditions:  . Thus, without loss of generality, the translation-invariant Gaussian kernel, which is the most widely used Mercer kernel, was employed when computing correntropy [A1, A2] ([30, 32] in revised manuscript). As we know, the kernel function of correntropy is almost Gaussian kernel. Although References [A3, A4] proposed mixture correntropy and multi-kernel correntropy for robust learning, the essence of these methods is to combine several Gaussian functions to construct the kernel function. Therefore, based on common sense and

above experiences, we select Gaussian kernel as the kernel function.

[A1] I. Santamaria, P. P. Pokharel, J. C. Principe, Generalized correlation function: Definition, properties, and application to blind equalization. IEEE Trans. Signal Process. 54(6), 2187-2197 (2006).
[A2] P. P. Pokharel, W Liu, J. C. Principe, A low complexity robust detector in impulsive noise. Signal Process. 89(10), 1902–1909 (2009).
[A3] B. Chen, X. Wang, N. Lu, S. Wang, J. Cao, J. Qin, Mixture correntropy for robust learning, Pattern Recognit., 79, 318–327 (2018).
[A4] B. Chen, X. Wang, Z. Yuan, P. Ren, J. Qin, Multi-kernel correntropy for robust learning, 2019.[Online]. Availabvle: http://arxiv.org/abs/1905.10115v1.

Comments No.2: In the manuscript, the symbol of 'sigma' is used to refer sigmoid function and Gaussian kernel. If possible try to differentiate between the two.

Answer:
Sorry for our carelessness. We have modified the symbol that refers to the sigmoid function to . Please refer to Eqs. (5), (6) and (8) with red font in the revised manuscript.

Comments No.3: In addition to PESQ and STOI, the ESTOI value could also be computed to reveal the spectral deviations and improvement after mapping the noisy speech to the approximate clean speech.

Answer:
According to the reviewer’s advice, the ESTOI metric has been added when evaluating the performance of speech enhancement. Please refer to Fig. 6 and the second paragraph of Subsection 3.3.2 1) in the revised manuscript.

Comments No.4: Overall the manuscript need to go through proofreading to improve the grammatical errors, typos, and English. Typo in Fig. 2 must be corrected.

Answer:
Thanks for the reviewer’s advice. We have checked the grammatical errors, typos and English throughout the manuscript carefully, and further improved English quality of our manuscript. Typos in Fig. 2 (Fig. 3 in revised manuscript) have been corrected, please refer to page 9.

Comments No.5: Use of abbreviations are not consistent. If you are using abbreviations, define it the first time you are using and do not expand it next time you use it in the manuscript.

Answer:
Thanks for the reviewer’s advice. We have carefully checked the use of abbreviations throughout the manuscript.

Comments No.6: Some abbreviations are mis-handled, the authors must avoid such mistake, for e.g., SFTF for STFT and MMC for MCC.

Answer:
Sorry for our carelessness. We have carefully corrected the spelling mistakes of the abbreviations throughout the manuscript. The “SFTF” and “MMC” have been modified to “STFT” and “MCC” with red font in the revised manuscript.

Comments No.7: In subsection 2.3 please provide relevant references for which you are referring to the recent studies.

Answer:
According to the reviewer’s advice, we have cited the relevant references in the revised manuscript. Please refer to the references [18, 37] in subsection 2.3.

Reviewer3:
Comments No.1: The authors should clearly state how their current proposal is different from the previous work [27], which is also based on multi-objective based multi-channel speech enhancement. A point-wise contribution of the current work is required apart from showing the differences from the previous work.

Answer:
Thanks for the reviewer's suggestion. Indeed, the current proposal is an extension of the previous work [27], but these two methods have three main differences. First, the loss function is different. This work proposed a model based on correntropy loss function, while the model in Ref. [27] used MSE as loss function. Second, the datasets used for training and testing are different. In this work, although clean and noise speech are the same as Ref. [27], the way of constructing noisy speech is totally different. In addition, the acoustic conditions have also extended to more severe conditions. As shown in Table I and Table II, the RT60 are set to {0.3s, 0.4s, 0.6s, 0.8s, 0.9s} during training, and {0.45s, 0.7s} during testing. But in Ref. [27], the maximum of RT60 was only set to 0.6s and 0.57s during training and testing, respectively. Third, the model structure is different. This work applied a BiLSTM to learn multi-channel multi-objective function simultaneously, while Ref. [27] employed a BiLSTM with shared weights, which processes the input of each channel separately. Besides, after the LPS fusion layer, a correntropy based loss function is also added to further boost the learning capability of the proposed model.

Therefore, the main contributions of this manuscript are summarized as follows: 1) A correntropy based multi-objective multi-channel speech enhancement model is proposed, i.e. correntropy is employed to construct a multi-objective loss function to optimize the model; 2) The model structure is designed more reasonable, and displays a superior enhanced performance in untrained acoustic conditions; 3) More severe acoustic conditions have been considered during training to improve the generalization of the proposed model.

According to the reviewer's comment, we have added the above-mentioned point-wise contribution of the current work in the revised manuscript. Please refer to the penultimate paragraph in Section 1.

Comments No.2: The authors do not show any illustration of noisy speech considered and the enhanced speech in terms of waveforms or spectrograms, which would have been useful for the readers. Additionally, if they could release some of the noisy speech examples and corresponding enhanced samples for the readers.

Answer:
According to the reviewer's advice, the waveforms and spectrograms, and the corresponding description of analyses have been added to further demonstrate the effectiveness of the proposed method. Please refer to Fig. 8 and the last paragraph of Subsection 3.3.2 2) in the revised manuscript. Besides, the noisy speech examples and corresponding enhanced speech samples are released in "http://faculty.dlut.edu.cn/chenzhe/zh_CN/jxzy/745520/content/3834.htm#jxzy".

Comments No.3: The authors have used TIMIT database for studies with additional noises. It would have been nice if the methods could have been validated using DNS challenge data, which is recently used for speech enhancement research.

Answer:
According to the reviewer's suggestion, the wide band track of DNS challenge data have been used to validated the generalization of the proposed method. Please refer to the Subsection 3.3.5 in the revised manuscript.

Comments No.4: The authors do not compare their methods with some of the latest state-of-the-art methods such as MetricGAN (https://github.com/speechbrain/speechbrain). It is always good to compare the proposed method with some latest methods.

Answer:
Thanks for the reviewer’s advice. MetricGAN is designed to solve the discriminator-evaluation mismatch (DEM) problem in GAN, and presents superior performance in speech enhancement task. However, it is a single-channel based method, and we think it is unfair to compare it with our multi-channel based method. Therefore, a latest multi-channel method named multi-resolution convolutional auto-encoders (MRCAE) [A1] ([46] in revised manuscript) is employed for comparison, and the results have been added in all the experiments. Please refer to the Tables 6-11 and Figs. 5-11 and the corresponding descriptions of Subsections 3.2, 3.3 in the revised manuscript.

[A1] E. M. Grais, D. Ward, M. D. Plumbley, Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders, in: European Signal Processing Conference (EUSIPCO), pp. 1577-1581 (2018).

Comments No.5: In Section 3.3.1, it is mentioned "From Table 3, when the value of \sigma 1.0, 1.3, 1.4 and 1.5, the performance of the model tends to be the best." However, the reviewer finds that when \sigma =1.0, it is not the best. The PESQ value at 1.8 is better than 1.0. Thus, it is a bit strange to consider \sigma=1.0 for the next study shown in Table 4 and then to report it as the best case.

Answer:
Sorry for our unclear description. Table 3 shows the average PESQ results with different kernel sizes . When the value of is 1.0, 1.3, 1.4 and 1.5, the corresponding PESQ is 2.219, 2.218, 2.218, and 2.218, respectively. However, when the value of is 1.8, the PESQ is only 2.217. Thus, we mentioned that “From Table 3, when the value of 1.0, 1.3, 1.4 and 1.5, the performance of the model tends to be the best.”, and selected 1.0 as the value of in the following experiments.

Comments No.6: In Section 3.3.2, machinegun noise is used for study. Instead of machinegun noise, some commonly occurring noise like vehicle noise would be nice to observe.

Answer:
In our manuscript, three noise datasets (NOISEX-92, Aurora2 and DNS) are employed to evaluate the enhanced performance of the proposed method. They contain the commonly occurring noise, such as vehicle noise. From the results of overall performance in Table 7 and Fig. 7, untrained noise types in Table 9 and Fig. 10 and DNS dataset in Table 10 and Fig. 11, it can be observed that the proposed method has good robustness to commonly occurring noises. Besides, machinegun noise is separately used only to study the advantage of the proposed method under impulse noise. Therefore, considering the page limitation, we did not conduct another vehicle noise experiment in the revised manuscript.

Comments No.7: The reviewer guessed in page 13, overall performance evaluation should be a new subsection.

Answer:
Yes, “overall performance evaluation” is a new subsection. To make it clearer, we have added a label for this subsection. Please refer to Subsection 3.2.2 2) in the revised manuscript.

Comments No.8: In Table 8, why did C-IRM perform the best for low SNR input?

Answer:
Table 8 shows the average PESQ results with untrained noise types at different SNRs. It can be seen that the waveform-based method MRCAE has the best performance. However, for the mapping- and masking- based methods, C-IRM performs the best for low SNR input, and the reasons are summarized as follows. First, when the input SNR is low, the clean speech are almost submerged in noises, which makes the predicted errors of multi-objective based methods lager than those of single-objective based methods. Second, compared with LPS feature, IRM estimates the probability of speech at each frequency, and has the capability to retain more useful information of clean speech and reduce speech distortion in time domain. Third, the loss function based on

| | |
|---|---|
| | correntropy is more insensitivity to outliers (impulse noise) than those based on MSE. Based on above reasons, C-IRM performs the best among the mapping- and masking-based methods at low SNR input.<br><br>Comments No.9: There are several typos in the paper. Authors should read carefully and revise. Some common typos are:<br>(i) page 2, level differences (ILD)) => level differences (ILD)<br>(ii) page 3, we extend our previous work [27] to proposed a ... => we extend our previous work [27] to propose a ....<br>(iii) page 5, Generally, its value is range from ... => Generally, its value ranges from ...<br><br>Answer:<br>Thanks for the reviewer's kind comment. We have carefully corrected the careless typos throughout this manuscript. The "level differences (ILD))" has been modified to "level differences (ILD)" , "we extend our previous work [27] to proposed a ..." has been modified to "we extend our previous work [27] to propose a…" and "Generally, its value is range from …" has been modified to "Generally, its value ranges from ...<br>" with red font in the revised manuscript. Besides, other several typos have also been checked carefully and corrected totally in the revised manuscript. |

# Correntropy based Multi-objective Multi-channel Speech Enhancement

Xingyue Cui[1], Zhe Chen[1,*], Fuliang Yin[1], Xianfa Xu[1]

*School of Information and Communication Engineering, Dalian University of Technology, Dalian 116023, China*

## Abstract

Although deep learning-based methods have greatly advanced the speech enhancement, their performance are intensively degraded under the non-Gaussian noises. To combat the problem, a correntropy based multi-objective multi-channel speech enhancement method is proposed. First, the log-power spectra (LPS) of multichannel noisy speech are feed to the bidirectional long short-term memory (BiLSTM) network with the aim of predicting the intermediate ideal ratio mask (IRM) and LPS of clean speech in each channel. Then, the intermediate LPS and IRM features obtained from each channel are integrated into a single-channel LPS and IRM feature by using a fusion layer, respectively. Next, the two single-channel features are fused into a single-channel LPS and feed to the deep neural network to learn the relationship between the predicted LPS and the clean speech LPS. Finally, a loss function for network training is constructed by correntropy to improve the performance of overall network. Experimental results show that the proposed method achieves significant improvements in suppressing non-Gaussian noises and reverberations, and has good robustness to different noises, signal-noise-ratios (SNRs) and source-array distances.

*Keywords:* Speech enhancement, microphone array, correntropy, neural networks, masking

## 1. Introduction

When acquiring speech signal in the real-world environment, it is inevitably suffering from noise and reverberation, and such nonideal signal does have a significant impact on the performance of following speech-related tasks. To improve the performance of speech processing tasks, such as speech recognition accuracy [1] and communication quality [2], some speech enhancement (SE) algorithms [3] were proposed to extract the desired speech from noisy-reverberation speech. Generally, speech enhancement methods can be divided into two categories. The first category uses a single microphone (also called monaural), while the second category uses multiple microphones (also called multichannel) to perform speech enhancement.

For monaural based speech enhancement, many approaches have been developed in the past decades. Spectral subtraction [4], Wiener filtering [5], subspace method [6] and statistical model

*Corresponding author: Zhe Chen

*Email addresses:* xiechoah@mail.dlut.edu.cn (Xingyue Cui), zhechen@dlut.edu.cn (Zhe Chen), flyin@dlut.edu.cn (Fuliang Yin), xuxianfa@mail.dlut.edu.cn (Xianfa Xu)

algorithm [7] are all typical noise reduction methods. Weighted prediction error (WPE) [8] and inverse filter [9] are both widely used methods for dereverberation. However, these methods can only achieve either denoising or dereverberation, and the former is less robust to non-stationary noise, while the latter is difficult to fully eliminate the late reverberation. More recently, the supervised learning methods have attracted wide attention since they can adapt to different acoustic conditions. In [10], a single deep neural network (DNN) was employed to simultaneously perform denoising and dereverberation via spectral mapping. In [11], a time-frequency masking called complex ideal ratio mask (cIRM) is proposed by Williamson and Wang for DNN-based enhancement. In the following [12], a two-stage system was proposed, where two DNN-based subsystems are sequentially conducted to denoising and dereverberation, then these pre-trained DNNs were combined into a deeper network for joint training. Subsequently, Li et al. [13] proposed a multi-objective speech enhancement learning framework based on stacked and temporal convolutional neural network, which uses the log-power spectra (LPS), power function compression Mel-frequency cepstral coefficient and ideal ratio mask (IRM) as the target features. The above-mentioned methods are all based on single-channel time-frequency (T-F) information and have achieved good effects in speech enhancement.

Nowadays, with multiple microphones being equipped on modern devices, microphone array techniques have become attractive solutions for speech enhancement. Spatial filtering (or so-called beamforming) is one of the most typical techniques [14]. Its idea is to design a linear filter to enhance or maintain the signal from the target direction while attenuate the interferences from other ones. Another effective method is based on a coherence algorithm that calculates the correlation of two input signals to estimate a filter to weaken the interference components [15], [16]. Besides, post-filtering [17] is commonly used for further noise reduction. Usually, when speech covariance matrices and direction of the arrival can be accurately estimated, the above methods are capable of achieving good speech enhancement performance. However, due to the complex environmental conditions, especially non-stationary noise and room reverberation, the capability of these methods are fundamentally limited. Meanwhile, the number of microphones in the array is also one of the restrictions.

More Recently, deep learning methods have exhibited encouraging performance in multichannel SE tasks. Its main idea is to estimate the time-frequency masks [18] using multi-channel information. For binaural enhancement, Jiang et al. [19] used DNNs to estimate the ideal binary mask (IBM), where monaural feature, binaural features ( interaural time (ITD) and level differences (ILD) are used for network training. In subsequent studies [20, 21], multiple features, such as spatial feature, spectral feature and phase differences between channels, were exploited as the network input to train a DNN or deep auto-encoder (DAE) for speech enhancement. Following in [22], the binaural signals were combined into a monaural complex signal, a complex mask was estimated using the complex DNN and then applied to the monaural complex signal.

In addition, some other methods [23-29] that are not based on the position of speech source have been proposed as well. In [23], a bi-directional long short-term memory (BiLSTM) network was adopted to compute the spectral masks of speech and noise, which then utilized to calculate the corresponding cross-power spectral density (PSD) matrices for getting the beamformer coefficients. Following, an extended version [24] was developed, where the spatial covariance matrices and beamforming are estimated by complex-valued short-time Fourier transform coefficients, and the magnitude features were used for mask prediction. In recent contributions [25, 26], Chakrabarty proposed a speech enhancement approach based on convolutional recurrent neural network (CRNN), where the estimated T-F masking is either directly applied to the microphone signal or indirect compute the PSD matrices for beamformer. In [27], multiple features of each channel were estimated by using a sharing BiLSTM network, which then fused to obtain the desired single-channel signal. Moreover, in addition to masking- or feature mapping- based methods, some other approaches have also been proposed. In [28], Higuchi et al. derived a frame-by-frame update rule for mask-based minimum variance distortion less response (MVDR) beamformer, which is capable of obtaining the enhanced signals without a long delay. Following, Qi et al. [29] proposed a tensor-to-vector regression approach, which casts the conventional DNN based vector-to-vector regression formulation under a tensor-train network (TTN) framework to address the issue of input size explosion and hidden-layer size expansion.

Although existing methods show strength in suppressing many kinds of noises, they still face with challenges due to non-Gaussian noises, especially impulse noise. The reason lies that most methods are based on mean square error (MSE) criterion which would be fragile to outliers. Thus, another loss function inspired by the statistical measure called correntropy [30, 31, 32] is proposed to improve the robustness of non-Gaussian noises. Correntropy is a nonlinear and local similarity measure that shows the similarity between two random variables in a neighborhood of the joint space controlled by the kernel bandwidth. Compared with MSE, the main advantage of correntropy is its insensitivity to outliers (or impulsive noises), which indicates that it has more potential for robust feature learning. Some related works have been done in recent years. Singh et al. [33, 34] proposed a correntropy based loss function for training network classifier, where the correntropy-loss (C-loss) displays superior robustness against the outliner and can approximate different norms (from L0 to L2) of data. Ref. [35] proposed a robust stacked autoencoder (RSAE) based on maximum correntropy criterion (MCC) to deal with the data containing non-Gaussian noises and outliers. Following, Chen et al. [36] proposed a model based on stacked auto-encoders (SAE) and correntropy-induced loss function (CLF). In this model, the reconstruction loss, the sparsity penalty term and the fine-tuning procedure were all built with CLF, and the results showed an obvious improvement of the model robustness when the data contains outliers and impulsive noise.

The motivation of this work are summarized in Table 1. In order to suppress impulse noise, we extend our previous work [27] to propose a correntropy based multi-objective multi-channel speech enhancement method in this paper. Specifically, a BiLSTM network is first trained to learn the mapping relationship between the multi-channel LPS of noisy speech and their corresponding clean LPS and log-ideal ratio mask (LIRM). Then, the intermediate outputs LPS and LIRM from multiple channels are separately fused into a single-channel LPS and a single-channel IRM. Finally, the two single-channel features are incorporated into a single-channel LPS and then feed to the fully connected layers for further predicting the LPS of clean speech. Different from conventional speech enhancement network that uses MSE as loss function, the objective function in the proposed method is built based on correntropy. Some experiments under different acoustic conditions verify the superior enhanced performance and the generalization of the proposed method.

The main contributions of this paper are summarized as follows: 1) A correntropy based multi-objective multi-channel speech enhancement model is proposed, i.e. correntropy is employed to

**Table 1** The motivation of correntropy based deep learning speech enhancement

| Num | Motivation |
|---|---|
| 1 | Compared with MSE, correntropy is insensitive to the outliers (or impulse noise). |
| 2 | Compared with traditional multi-channel speech enhancement methods, deep learning based methods do not explicitly require the position of speech source but still exhibit encouraging enhanced capability. |
| 3 | The performance of the existing deep learning based methods are severely degraded under the non-Gaussian noises, especially impulse noise. |

construct a multi-objective loss function to optimize the model; 2) The model structure is designed more reasonable, and displays a superior enhanced performance in untrained acoustic conditions; 3) More severe acoustic conditions have been considered during training to improve the generalization of the proposed model.

The rest of this paper is organized as follows. In Section 2, the formulation of the problem, the target features and related networks involved in this framework are introduced. Then, a detailed description of the correntropy based multi-objective multi-channel network is presented. Following, the experimental evaluations of the proposed method are provided in Section 3. Finally, Section 4 concludes the paper.

## 2. Multi-channel Speech Enhancement based on Correntropy

In the following subsections, the noisy-reverberant signal model for microphone array speech enhancement is first introduced. Then, the related features (LPS and IRM), the architecture of long-short term memory (LSTM) and the definition of correntropy are elaborated in details. Finally, the process of correntropy based multi-objective multi-channel speech enhancement is described, including the overall learning framework and the procedure of network training.

### *2.1. Problem Formulation*

Consider a reverberant and noisy environment, the signals received by $M$ microphones are typically modeled as

$$\mathbf{X}(n,k) = \mathbf{H}(k)\,\mathbf{S}(n,k) + \mathbf{V_d}(n,k) + \mathbf{V}(n,k) \tag{1}$$

in the frequency domain. Here, $\mathbf{X}(n,k)$, $\mathbf{S}(n,k)$, $\mathbf{V_d}(n,k)$ and $\mathbf{V}(n,k)$ are the $M$-dimensional vectors that denote the short-time Fourier transform (STFT) of the received signals, clean signals, environmental noises and spatially uncorrelated microphone self-noises at the $n$-th time frame and the $k$-th frequency bin, i.e. $\mathbf{X}(n,k) = [X_1(n,k), X_2(n,k), \cdots, X_M(n,k)]^T$, and $\mathbf{H}(k)$ denotes the STFT of the acoustic transfer function. In this paper, the goal of speech enhancement is to find a function that maps the noisy-reverberant observation $\mathbf{X}(n,k)$ to the clean speech component that approximates $\mathbf{S}(n,k)$ as close as possible.

### 2.2. Spectral Features

In recent studies, the speech log power spectrum (LPS) is usually preferred as the learning targets because of the broad dynamic range of spectral magnitude. Given the received $M$-dimensional time domain signals, the corresponding frame length and frame shift are set as $N$ and $N/2$ samples. Then, a Fourier transform is applied to each overlapping windowed frame, and each frame can be described using a vector $\mathbf{x}(n)$ as

$$\mathbf{x}(n) = [\mathbf{X}(n, 0), \mathbf{X}(n, 1), \cdots, \mathbf{X}(n, N)]^{\mathrm{T}} \tag{2}$$

Since speech is correlated from frame to frame, we incorporate temporal dynamics by joining adjacent frames into a single feature vector. Therefore, the input features for network are extended as

$$\tilde{\mathbf{x}}(n) = [\mathbf{x}(n-l), \cdots, \mathbf{x}(n), \mathbf{x}(n+1), \cdots, \mathbf{x}(n+l)]^{\mathrm{T}} \tag{3}$$

where $l$ is the number of adjacent frames on each side, then the total number of involved frames is $2l + 1$. Besides, similar to the input of network, the desired output of network is also the log power spectra at the $n$-th frame, and all the input and output features are synchronously extracted from noisy and clean speeches to keep aligning on each frame.

### 2.3. Ideal ratio mask

Recent studies [18], [37] have shown that ratio mask targets are superior to other ones in terms of the objective intelligibility and quality metrics. Thus, one of the mask targets called ideal ratio mask (IRM) [37] is employed as another output feature in the proposed network. IRM is a kind of soft mask to measure the presence of speech in a T-F unit. Generally, its value ranges from zero to one. In this work, after considering the statistical independence between clean speech and noise, the IRM of each T-F bin is defined as

$$IRM(n,k) = \frac{|\mathrm{S}(n,k)|}{|X(n,k)|} \tag{4}$$

where $S(n,k)$ and $X(n,k)$ denote the clean- and noisy- spectral magnitudes at the $k$-th frequency bin in the $n$-th frame, respectively. From Eq. (4), it can be seen that unlike the convention IRM, the values of IRM defined here may greater than one. Thus, in order to avoid unwanted amplification of noise components in the signal due to estimation errors and obtain better numerical stability in backpropagation training, the values are saturated to one. The whole computation process of IRM is shown in Fig .1.

In addition, since IRM is adopted as an intermediate output to merge with another intermediate output LPS, a logarithmic operation is performed on all IRM for convenience, denoted as LIRM.

### 2.4. Long-short Term Memory Network

Recurrent neural network (RNN) displayed excellent performance in solving time series learning problems by adding the connections between output and hidden layers. However, it has a main obstacle as gradient vanishing problem [38] which limits the capability of learning long-range context dependencies. To address this problem, an improved RNN version is introduced as LSTM

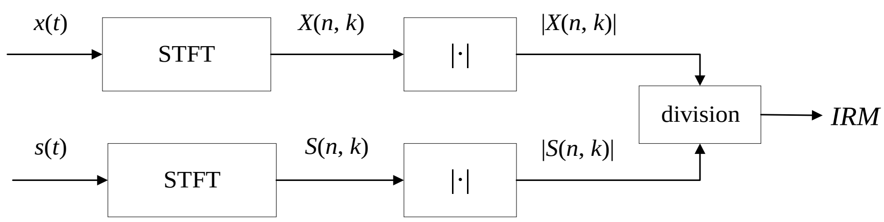



**Fig. 1.** The computation process of IRM.

[39], which defines memory cell and several gates to regulate the information flow. Fig. 2 illustrates the structure of LSTM memory block. It can be seen that each LSTM block comprises three gates, input gate $\mathbf{i}_t$, forget gate $\mathbf{f}_t$ and output gate $\mathbf{o}_t$, and the main implementation can be described as follows

$$\mathbf{i}_t = \kappa\left(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i\right) \tag{5}$$

$$\mathbf{f}_t = \kappa\left(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f\right) \tag{6}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot g\left(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c\right) \tag{7}$$

$$\mathbf{o}_t = \kappa\left(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o\right) \tag{8}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot g\left(\mathbf{c}_t\right) \tag{9}$$

where $\mathbf{x}_t$ and $\mathbf{h}_t$ are the input and output vectors at the time frame $t$, $\mathbf{c}_t$ is the memory cell state, $\mathbf{W}(\mathbf{W}_{xi}, \mathbf{W}_{hi}, \mathbf{W}_{xf}, \mathbf{W}_{hf}, \mathbf{W}_{xc}, \mathbf{W}_{hc}, \mathbf{W}_{xo}, \mathbf{W}_{ho})$ are the weight matrices that need to be learned during training and $\mathbf{b}(\mathbf{b}_i, \mathbf{b}_f, \mathbf{b}_c, \mathbf{b}_o)$ are the bias term of the corresponding gates. $\odot$ is the element-wise vector product, $\kappa(\cdot)$ and $g(\cdot)$ are the well-known sigmoid function and the hyperbolic tangent (tanh) function, which are usually used as gate activation function and input and output activation function.

In this work, the bidirectional long short-term memory (BiLSTM) network [40] is employed. A BiLSTM network has recurrent connections in both forward and backward directions, and this structure provides the output layer with complete past and future context in the input sequence. Thus, the BiLSTM can make full use of the temporal information and more suitable for speech enhancement.

### *2.5. Correntropy and maximum correntropy criterion*

For neural network training, one of the key issues is to select an appropriate loss function to measure the error between the output and the target so that the performance of whole network is the most optimal. As a classical second-order statistic, mean square error (MSE) is preferred as a loss function in many supervised learning networks. However, it is sensitive to non-Gaussian noises and outliers so that the capability of feature learning would be fragile when the input is high noisy-reverberation data. Hence, in this paper, correntropy [30] is employed as the loss function to optimize network performance.

Given two random variables $A$ and $B$, the cross correntropy (CC) (usually simply correntropy) is defined as [31]

$$V_\sigma(A, B) = \mathrm{E}\left[\kappa_\sigma(A - B)\right] \tag{10}$$

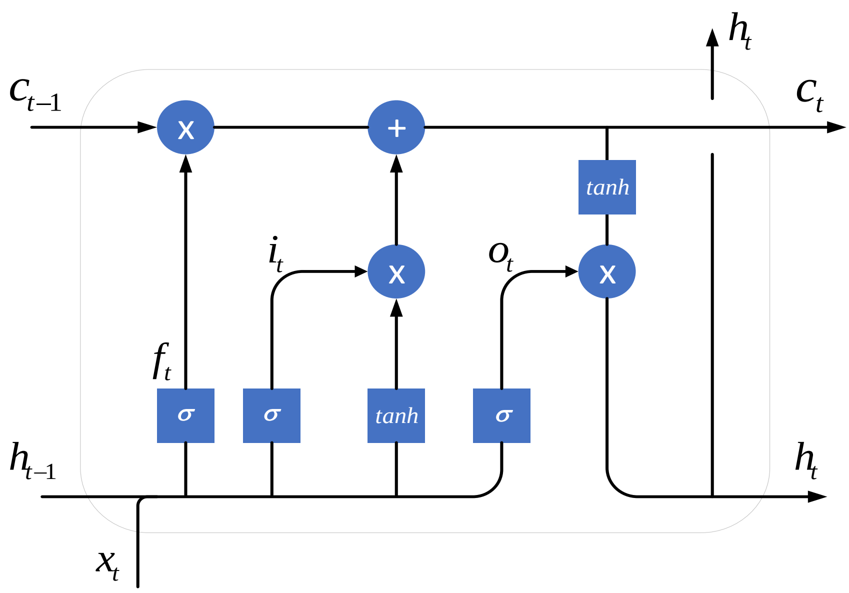



**Fig. 2.** An illustration of the LSTM block.

where E is the expectation operator, $\kappa_\sigma(\cdot)$ is the Gaussian kernel with $\sigma$ is the kernel size:

$$\kappa_\sigma(a - a_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|a - a_i\|^2}{2\sigma^2}\right) \tag{11}$$

Usually, due to the fact that the joint probability density function (PDF) is difficult to obtain and only a finite number of data $\{(a_i, b_i)\}_{i=1}^{N}$ are available, the sample estimation of the cross correntropy can be obtained by

$$\hat{V}_{N,\sigma}(A, B) = \frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(a_i - b_i) \tag{12}$$

Correntropy is a measure of the similarity between two random variables in a small neighborhood determined by the kernel size. Compared to MSE, these two measures have both similarities and significant differences. Generally, correntropy and MSE can be used as a certain optimization criterion as well as a measurement of similarity between random variables $A$ and $B$. However, MSE is a global statistical function, which is more suitable for Gaussian distribution. When the samples far away from the distribution center, MSE will significantly amplify the errors, thus, it cannot be optimal in the case of asymmetric error distribution, non-zero center and outliers. In contrast, correntropy is a local criterion of similarity, and has the capability to suppress the outliers because of the nonlinear kernel function, such as Gaussian kernel function. According to the characteristic of Gaussian function, the lager the outliers, the better the suppression. Thus, it is very suitable for cases when the error samples are non-Gaussian with large outliers [31, 34].

As a result, correntropy of the error is employed as a cost function in our network training, the goal of which is to maximize the similarity between the predicted output and the true values in the sense of correntropy. This is also called maximum correntropy criterion (MCC). In addition, the MCC is based on the property that maximizing the correntropy is equivalent to minimizing another metric named correntropy induced metric (CIM), i.e. the smaller the CIM, the larger the MCC, and the higher similarity between two features. In [31], the CIM is defined as

$$\mathrm{CIM}(X, Y) = (\kappa_\sigma(0) - V_\sigma(A, B))^{1/2} \tag{13}$$

which can be seen as a linear operation of correntropy and a constant. Therefore, in this paper, to facilitate using the MCC for network training, CIM is adopted as the loss function.

### 2.6. Multi-objective Multi-channel Speech Enhancement Scheme based on correntropy

Fig. 3 presents the architecture of the proposed microphone array speech enhancement network, which mainly consists of two stages: (1) multi-objective based multi-channel feature learning; (2) fusion and output. Overall, by using an array with $M (M > 2)$ microphones, a multi-channel signal can be collected from reverberant-noisy speech. Then, LPS features of these signals are extracted and used as the input of the network to simultaneously predicted the intermediate outputs (LPS and LIRM) of multiple channels. Subsequently, these multi-channel intermediate outputs are fused into a single-feature LPS and finally acquired the output of the proposed network. More specifically, in the first stage, a BiLSTM module composed of two layers BiLSTM is employed for the sake of learning the temporal spatial characteristics between the multi-channel LPS input and their corresponding clean LPS and LIRM. Following is the fully connected module, which $M$ Dense layers are first performed as $M$ output channels, and then each output channel predicts intermediate LPS and LIRM output by using two independent Dense layers. In the second stage, fusion layers are separately operated to cope with the directly predicted LPS and the indirect LPS obtained from the LIRM estimation and then fused these two to form the final single-channel LPS feature. Here, the average sum operation is used for all the fusion layers

$$\hat{Y}_d^f(n, k) = \frac{1}{M} \sum_{m=1}^{M} \hat{Y}_m(n, k) \tag{14}$$

$$\hat{Y}_i^f(n, k) = \frac{1}{M} \sum_{m=1}^{M} [Y_m(n, k) + I_m(n, k)] \tag{15}$$

where, $\hat{Y}_m(n, k)$, $Y_m(n, k)$ and $I_m(n, k)$ are the directly predicted clean LPS, the corresponding noisy LPS input and mask from channel $m$, $Y_m(n, k) + I_m(n, k)$ is the masking-based indirect LPS feature. Ultimately, the ensemble results $\hat{Y}_d^f(n, k)$ and $\hat{Y}_i^f(n, k)$ are merged into a single channel output and feed to the fully connected layers to get the desired clean LPS. It should be pointed out that because the fusion LPS is used as secondary prediction target, otherwise, all the $2M$ channel features can be directly fused into a single one.

As for the network configuration, 1024 hidden units are adopted in every BiLSTM layer. For fully connected module in each channel, 512 hidden units are first used and then 129 units are introduced with the aim of predicting the intermediate LPS and LIRM. Meanwhile, after the fusion layers, 512 hidden units are used for two DNN layers, and 129 units are employed in output layer to estimate the clean LPS. Moreover, tanh and hard-sigmoid are used as the activation function

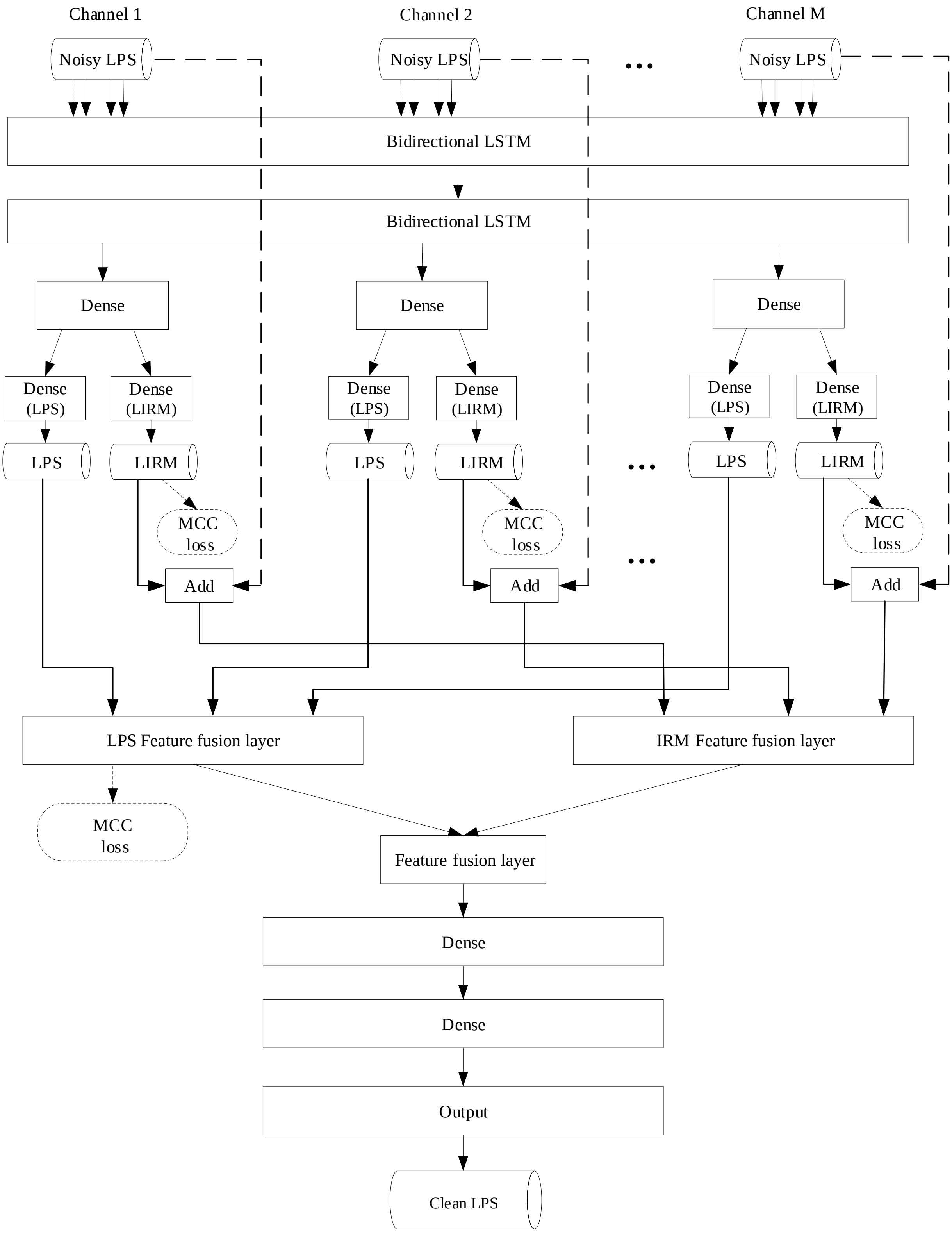



**Fig. 3.** Block diagram of correntropy based multi-objective multi-channel speech enhancement scheme.

and recurrent activation function for recurrent layers, respectively. All the DNNs are trained with the rectified linear unit (ReLU) and the linear unit is utilized for the estimation of LIRM and LPS.

Since the network contains multiple output layers with multiple targets to be predicted, the following multi-objective MCC (minimum CIM) is constructed as the loss function

$$L = \frac{1}{T}\left[\sum_{m=1}^{M} \mathrm{CIM}(\hat{I}_m, I_m) + \mathrm{CIM}(\hat{S}_{LPS}, S) + \mathrm{CIM}(\hat{S}, S)\right] \tag{16}$$

where $\hat{I}_m$ and $I_m$ indicate the predicted LIRM and its corresponding clean IRM in the $m$-th channel; Similarly, $\hat{S}_{LPS}$, $\hat{S}$ and $S$ indicate the intermediate fused LPS, the final predicted LPS and their corresponding clean LPS; and $T$ is the mini-batch size. Here, all the time and frequency indexes are omitted for more intuitive representation. Besides, Adam is utilized as the optimizer to train the network.

As described above, the exploitation of multiple objectives and multiple channels can fully utilize the complementarity of speech features and increase the diversity of training data [41]. Moreover, the fusion layer can alleviate the problem of overestimation or underestimation of the enhanced spectrum. Besides, correntropy based loss function is insensitive to outliers and more robust to non-Gaussian noises. In summary, the proposed scheme has a better potential for denoising and dereverberation.

## 3. Experimental Results and Discussions

In this section, several comparative experiments are presented to evaluate the performance of the proposed method. First, the data generation and experimental setup are introduced. Then, the generalization capability of the proposed scheme is evaluated under different acoustic conditions (RIRs, SNRs and noise types). Finally, the source-array distance is also discussed to further verify the robustness of the algorithm.

### *3.1. Dataset and Experimental Setup*

As shown in Fig. 4, a square array with $M = 4$ microphones is considered in this paper, and the distance from each microphone to the center of the square is 10 cm. For both training and testing, since the proposed method is aimed to be independent of the location of the speech source, the complete angular range of the array center is discretized with a step 5 degree, and 72 different angular positions of the speech source are obtained. Specifically, to generate the training data, five rooms with different acoustic conditions are employed, as shown in Table 2. For each array position and source-array position in each room, 8 speech sentences from TIMIT training set [42], are used for each angular position, i.e. convolved with the room impulse responses (RIRs) corresponding to the specific setup. Here, the RIRs that simulate different acoustics conditions are generated by the RIR generator [43]. Therefore, a total of $8 \times 72 = 576$ clean speech sentences are available in each array position during the training phase.

As for diffuse noise used in training phase, the NOISEX-92 database [44] which contains 15 types of noises are considered. By randomly choosing one of 15 types, the noisy speech

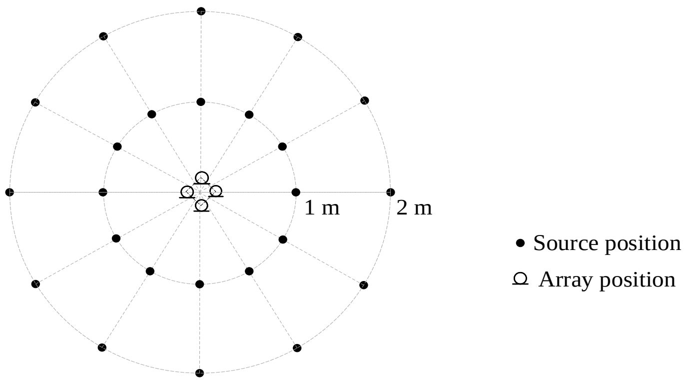



**Fig. 4.** Geometric setup for data generation.

**Table 2** Acoustic Conditions in Five Different Rooms for Training

| Parameters | Configuration on training data |
|---|---|
| Speech | TIMIT training set |
| Noise | NOISEX-92 |
| Room | R1:$(5\times4\times3)m^3$, R2:$(6\times5\times3)m^3$, R3:$(10\times5\times3)m^3$, R4:$(10\times8\times3)m^3$, R5:$(11\times8\times3)m^3$ |
| Array positions | 5 different positions in each room |
| Source-array distance | 1m, 2m |
| RT60 | R1: 0.3s, R2: 0.4s, R3: 0.6s, R4: 0.8s, R5: 0.9s |
| SNR | Diffuse: -5dB to 10dB, Spatially white Gaussian: 20dB |

are generated with signal-noise-ratio (SNR) in -5dB, 0dB, 5dB and 10dB. Meanwhile, spatially uncorrelated white Gaussian noise with 20dB SNR is also added as microphone self-noise.

For testing, different from training configuration, two mismatched acoustic conditions with arbitrary array positions as well as different source array distances are taken into account, as shown in Table 3. Moreover, for each array setup in each room, five sentences from TIMIT testset are used for each angular position of the source, thus, a total of 360 different sentences are available in each array position during the testing phase.

**Table 3** Acoustic Conditions in Five Different Rooms for Test

| Parameters | Configuration on test data |
|---|---|
| Speech | TIMIT test set |
| Noise | Aurora2 |
| Room | R1:(8×5×3)m$^3$, R2:(9×7×3)m$^3$ |
| Array positions | 3 arbitrary positions in each room |
| Source-array distance | 1.2m, 1.7m |
| RT60 | R1: 0.45s, R2: 0.7s |
| SNR | Diffuse: -2dB to 7dB, Spatially white Gaussian: 20dB |

In the proposed network, both BiLSTM and Dense networks are initialized with random weights, the hidden units are 1024 and 512, respectively. During training, the mini-batch size is 1024, the learning rate is 0.001 in the first 10000 epoch and then decreased to 0.0001. Besides, for each channel, all the clean and noisy waveforms are resampled to 8KHz, the corresponding frame length and frame shift are set as 256 samples (i.e. 32 ms) and 128 samples, respectively. Then, the log-power spectra of each frame are extracted as input feature, and its vector dimension is 129. All the implementations are done in Keras [45].

### *3.2. Reference Methods and Evaluation Metrics*

For comparison purposes, we construct mapping-based, masking-based and waveform-based neural networks as baseline. For the mapping-based network, the log power spectra of clean speech are directly estimated from those of noisy reverberant speech. For the masking-based network, time-frequency masks (IRM) are predicted and then combined with noisy reverberant speech to obtained clean speech. For the waveform-based method, time-domain waveforms of clean speech are directly estimated from corresponding noisy speech, without feature extraction. During training in mapping- and masking-based networks, MSE and correntropy are separately adopted as loss function to evaluate its effect on network performance. For convenience, in the following experiments, MSE-based methods are referred as **M-LPS** and **M-IRM** while correntropy-based methods are referred as **C-LPS** and **C-IRM**, respectively. Moreover, to make a fair comparison, the structure of mapping- and masking-based methods are basically same as that of the proposed method, where two-layer BiLSTM and two-layer Dense are employed, followed by a fusion block which combines the intermediate multi-channel output into single one, and lastly a Dense module is incorporated to acquire the final output. As for the waveform-based method, a multi-channel input and output network, called multi-resolution convolutional auto-encoder (MRCAE), is implemented according to [46], where the encoder and decoder are composed of two convolutional layers with sets of filters and two transposed convolution layers with sets of filters, respectively, and another transposed convolution layer is used for the final output. In addition, all the parameters of comparison networks are set same as the proposed network, and the number of the filters for each convolutional layer in MRCAE is set as Ref. [46]

The enhanced speech signals from each approach are evaluated with three objective metrics, namely perceptual evaluation of speech quality (PESQ) [47], short-time objective intelligibility (STOI) score [48] and extended STOI (ESTOI) [49]. PESQ is computed by comparing the enhanced speech with the corresponding clean speech, producing scores in range [-0.5, 4.5] where a higher score indicates a better quality. STOI measures speech intelligibility by computing the correlation of short-time temporal envelopes between clean and enhanced speech, resulting in scores in the range of [0, 1] where a higher score indicates a better intelligibility. In addition, PESQ scores have been shown to be highly correlated to human speech quality, while STOI and ESTOI scores have high correlation with human speech intelligibility.

### 3.3. Experiments with different acoustic conditions

In this section, a experiment is first designed to select a suitable value of kernel size for correntropy. Then, the generalization capability of the proposed method is explored by introducing two different acoustic environments in the following experiments. Besides, all the parameters configurations and comparison methods are depicted in Subsection 3.1 and 3.2, and the array positions showed in Table 3 are randomly selected in each room.

#### 3.3.1. Selection of kernel size $\sigma$

Since the correntropy is based on a Gaussian kernel function, the choice of the kernel size should always be relative to the dynamic range of the variable or signal. In other words, a suitable value of $\sigma$ is crucial to the performance of the network. However, speech enhancement is a regression task, the error between true values and predicted values are not fixed in a range like the classification tasks, thus, when using the correntropy as loss function, the kernel size $\sigma$ has to be chosen through the experimental results.

Generally speaking, for denoising and dereverberation, we are more concerned about whether the proposed method has better adaptability under different acoustic environments (Rooms). Therefore, this paper gives priority to the enhancement results of the C-LPS+IRM under different RIR conditions when choosing the kernel size. Table 4 lists the enhancement performance of the proposed C-LPS+IRM with different kernel size $\sigma$ under different RIRs. From Table 4, when the value of $\sigma$ is 1.0, 1.3, 1.4 and 1.5, the performance of the model tends to be the best. However, by looking in Table 5, although the PESQ results are similar when a = 1.0, 1.3, 1.4 and 1.5, more attention is paid to low SNR conditions for the overall intelligence of speech. Thus, $\sigma$= 1.0 is the most suitable kernel size and is fixed in the following experiments.

#### 3.3.2. Generalization to different RIRs

1) Generalization to impulse noise: To evaluate the robustness of correntropy based model against impulse noise, a typical impulse noise machinegun is tested in this experiment. As shown in Table 3, different RIRs and testing speech are employed, and the array positions in each room are randomly selected as well. Table 6 and Fig. 5 display the PESQ and STOI results under different

**Table 4** Average PESQ with different kernel size $\sigma$.

| $\sigma$ | PESQ | $\sigma$ | PESQ | $\sigma$ | PESQ |
|---|---|---|---|---|---|
| 0.5 | 2.212 | 1.2 | 2.215 | 1.7 | 2.215 |
| 0.8 | 2.215 | 1.3 | 2.218 | 1.8 | 2.217 |
| 0.9 | 2.212 | 1.4 | 2.218 | 1.9 | 2.215 |
| 1.0 | 2.219 | 1.5 | 2.218 | 2.0 | 2.216 |
| 1.1 | 2.215 | 1.6 | 2.217 | 2.5 | 2.215 |

**Table 5** PESQ results at different input SNRs with different kernel size $\sigma$.

| SNR(dB) | -5 | 0 | 5 | 10 |
|---|---|---|---|---|
| $\sigma$= 1.0 | 1.937 | 2.166 | 2.334 | 2.439 |
| $\sigma$= 1.3 | 1.931 | 2.166 | 2.334 | 2.440 |
| $\sigma$= 1.4 | 1.932 | 2.169 | 2.330 | 2.439 |
| $\sigma$= 1.5 | 1.934 | 2.166 | 2.331 | 2.440 |

noisy-reverberant conditions by using different methods. Obviously, C-LPS+IRM has a great advantage in suppressing machinegun noise, especially in the case of low input SNRs. However, in turn, MRCAE has the worst performance. Specifically, for PESQ metric, the average enhanced performance of C-LPS+IRM outperforms that of the second best C-IRM with 0.02 improvement. In addition, for low input SNRs (-5dB and 0dB), the gaps between these two models achieve to 0.038 and 0.025, respectively. Furthermore, for STOI metric, C-LPS+IRM still exhibits superior enhanced performance in all noisy-reverberant conditions. M-LPS+IRM and C-IRM jointly rank the second best with a comparable performance.

In addition to PESQ and STOI, another metric called extended STOI (ESTOI) is also employed to further evaluate the proposed method. Different from STOI, ESTOI works for a larger range of input signals, and does not assume mutual independence between frequency bands. Fig. 6 presents the ESTOI results of machinegun noise among seven models. It can be observed that although the results of ESTOI are decreased compared with those of convention STOI, the trends of model performance are basically consistent with STOI. C-LPS+IRM ranks first with a larger advantage compared with the second best C-IRM and M-LPS+IRM. M-IRM has the third rank, and MRCAE has the worst performance. From above, it can be stated that compared with MSE based models, the proposed correntropy based model is more capable of dealing with impulse noise.

2) Overall performance evaluation: To analyze the proposed model more comprehensively, the whole 15 noise types from NOISEX-92 are tested among different models. Same as the above experiment, the configurations of acoustic conditions are set as Table 3. Table 7 and Fig. 7 list the average PESQ and STOI results among seven models under different RIRs with trained noise types at different SNRs. It can be observed that first, all the seven models have the capability to obtain clean speech from noisy-reverberation speech. Second, for the mapping- and masking-based models, the enhanced performance of correntropy based models (C-LPS+IRM, C-IRM and C-LPS) are generally superior to those of MSE based models (M-LPS+IRM, M-IRM and M-LPS) in terms of both PESQ and STOI improvements. Third, although the IRM-based models generally

**Table 6** PESQ results of machinegun noise for different models.

| | PESQ | | | |
|---|---|---|---|---|
| SNR(dB) | -5 | 0 | 5 | 10 |
| Unprocessed | 1.423 | 1.725 | 1.924 | 2.032 |
| M-LPS | 2.176 | 2.248 | 2.285 | 2.304 |
| C-LPS | 2.180 | 2.247 | 2.292 | 2.308 |
| M-IRM | 2.158 | 2.321 | 2.431 | 2.469 |
| C-IRM | 2.212 | 2.354 | 2.447 | 2.491 |
| MRCAE | 2.110 | 2.156 | 2.182 | 2.199 |
| M-LPS+IRM | 2.215 | 2.352 | 2.440 | 2.476 |
| C-LPS+IRM | **2.253** | **2.379** | **2.462** | **2.501** |

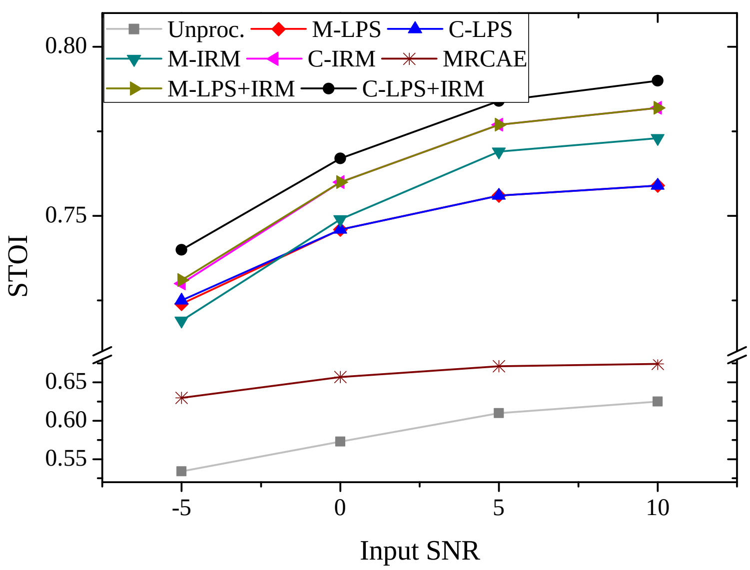



**Fig. 5.** STOI results of machinegun noise for different models.

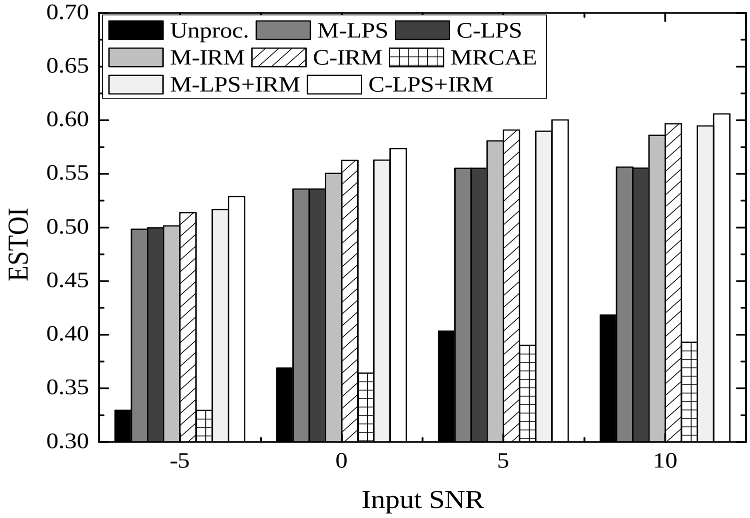



**Fig. 6.** ESTOI results of machinegun noise for different models.

perform better than LPS-based models, a further performance gain can be obtained by using the two features in combination, which implies that the LPS and IRM features are complementary to some extent. More specifically, PESQ results in Table 7 illustrate that C-LPS+IRM significantly outperforms other models with an average PESQ increment from 1.753 to 2.254 in Room 1, 1.682 to 2.183 in Room 2. Following are C-IRM and M-LPS+IRM with comparable performance, the improvements of PESQ are 0.485 and 0.48 in Room 1, 0.481 and 0.483 in Room 2, respectively. However, it can be noted that the performance of C-LPS is inferior to M-IRM, which indicates that a proper loss function does improve the performance of network, but still has a certain limitation. That is to say, many factors, such as feature selection, have effects on network performance. Moreover, STOI results in Fig. 7 show a similar trend to those of PESQ. C-LPS+IRM still ranks the first with the average improvement of 0.163 and 0.178 in Room 1 and Room 2. C-IRM and M-LPS+IRM have second and third rank with a gap of 0.5 and 0.7 percent, respectively.

To further illustrate the proposed method, an enhancement example is presented in Fig. 8. The waveforms and spectrograms of clean, noisy-reverberant and the corresponding enhanced speech obtained by MRCAE, C-LPS, M-IRM, C-IRM, M-LPS+IRM and C-IRM+IRM are shown in Fig. 8 (a), (b), (c), (d), (e), (f), (g) and (h), respectively. By comparing waveforms in Fig. 8(a) and Figs. 8(c)-(h), it is clear that the enhanced utterance from MRCAE is distorted notably, IRM-based methods lead to amplitude deviations, and LPS method loses speech segments at some frames. Moreover, by looking at the spectrograms, the utterance restored by C-LPS+IRM has a clearer structure, M-LPS+IRM is slightly inferior but still competitive. Following, C-IRM and C-LPS preserve excessive noise in the high-frequency components, and M-LPS exhibits some distortion in the low-frequency components. Besides, MRCAE fails to remove the background noise and reverberation.

**Table 7** Average PESQ results under untrained RIRs at different SNRs.

| | Room1 | | | | Room2 | | | |
|---|---|---|---|---|---|---|---|---|
| SNR(dB) | -5 | 0 | 5 | 10 | -5 | 0 | 5 | 10 |
| Unprocessed | 1.454 | 1.678 | 1.867 | 2.014 | 1.427 | 1.614 | 1.783 | 1.906 |
| M-LPS | 1.901 | 2.104 | 2.232 | 2.304 | 1.888 | 2.061 | 2.170 | 2.242 |
| C-LPS | 1.912 | 2.109 | 2.243 | 2.314 | 1.894 | 2.066 | 2.180 | 2.252 |
| M-IRM | 1.904 | 2.158 | 2.341 | 2.467 | 1.871 | 2.092 | 2.258 | 2.355 |
| C-IRM | 1.933 | 2.183 | 2.356 | 2.482 | 1.898 | 2.118 | 2.270 | 2.368 |
| MRCAE | 1.951 | 2.073 | 2.154 | 2.211 | 1.896 | 2.006 | 2.075 | 2.114 |
| M-LPS+IRM | 1.936 | 2.174 | 2.351 | 2.472 | 1.904 | 2.114 | 2.278 | 2.367 |
| C-LPS+IRM | **1.952** | **2.197** | **2.374** | **2.493** | **1.922** | **2.133** | **2.294** | **2.386** |

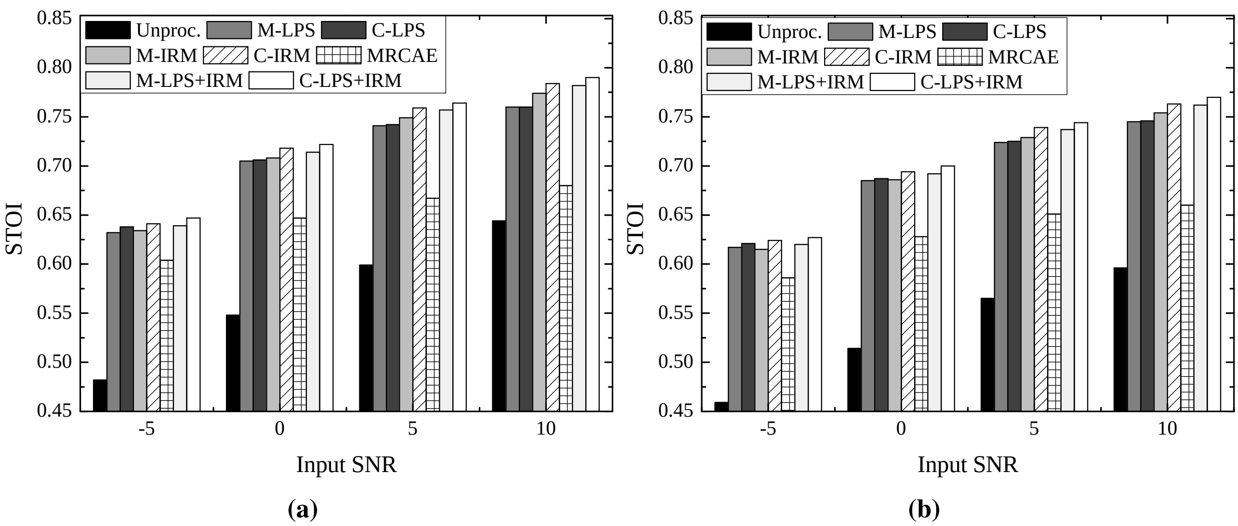



**Fig. 7.** Average STOI results in (a) Room 1 and (b) Room 2 at different SNRs.

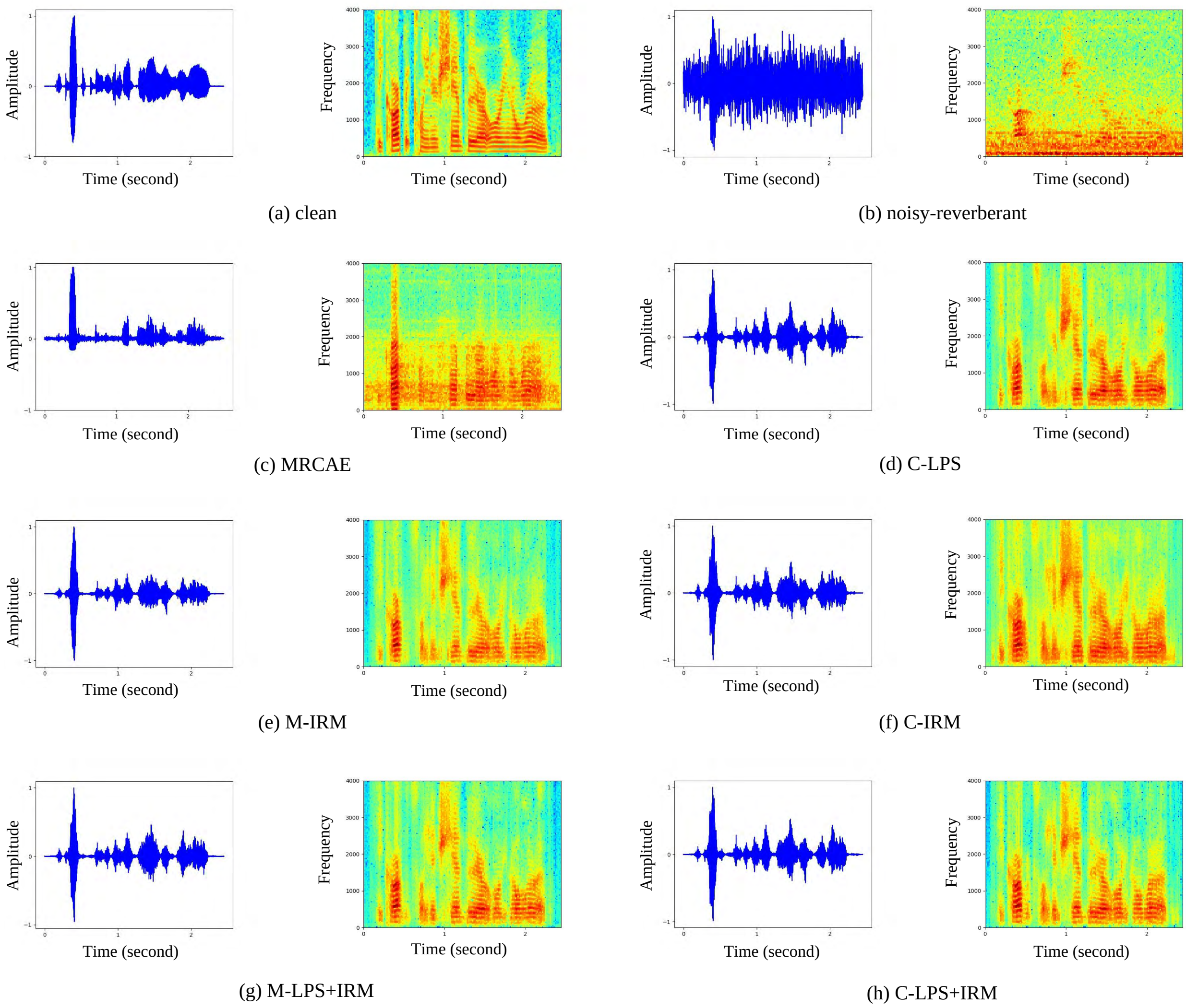



**Fig. 8.** Waveforms and spectrograms of an example utterance: (a) clean speech; (b) noisy-reverberant speech (m109 noise, SNR=-5dB, RT60=0.7s); (c) enhanced speech by MRCAE; (d) enhanced speech by C-LPS; (e) enhanced speech by M-IRM; (f) enhanced speech by C-IRM; (g) enhanced speech by M-LPS+IRM; (h) enhanced speech by C-LPS+IRM.

#### 3.3.3. Generalization to untrained SNRs

For the sake of illustrating the proposed C-LPS+IRM model is not sensitive to untrained SNRs, in this experiment, we choose -2dB and 7dB that are not involved in training as input SNRs. Table 8 and Fig. 9 present the average PESQ and STOI scores among six models with trained noise types at untrained SNRs. On the whole, it can be found that the model still attain excellent enhancement capability even facing with untrained SNRs, and are basically consistent with the results of trained SNRs shown in Table 7 and Fig. 7. For PESQ, C-LPS+IRM provides improvements of 0.518 and 0.487 over unprocessed speech in Room 1 and 2, ranks the first. Next, M-LPS+IRM has the second best with the 0.504 and 0.512 improvement in Room 1 and Room 2. For STOI, the proposed C-LPS+IRM still outperforms other models with nearly 2% improvements compared with noisy speech. Besides, the performance of M-LPS+IRM model are inferior to those of C-IRM model, which is slightly different from PESQ results. But taken together, they have a comparable performance, jointly rank the second best.

**Table 8** Average PESQ results at different untrained SNRs.

| | Room1 | | Room2 | |
|---|---|---|---|---|
| SNR(dB) | -2 | 7 | -2 | 7 |
| Unprocessed | 1.588 | 1.934 | 1.531 | 1.835 |
| M-LPS | 2.035 | 2.261 | 2.002 | 2.201 |
| C-LPS | 2.040 | 2.272 | 2.013 | 2.211 |
| M-IRM | 2.072 | 2.394 | 2.022 | 2.298 |
| C-IRM | 2.087 | 2.409 | 2.043 | 2.312 |
| MRCAE | 2.042 | 2.176 | 1.973 | 2.090 |
| M-LPS+IRM | 2.092 | 2.398 | 2.050 | 2.315 |
| C-LPS+IRM | **2.106** | **2.421** | **2.066** | **2.332** |

#### 3.3.4. Generalization to untrained noise types

In addition to the previous experiment that verified the adaptability of the proposed model to untrained SNRs, the robustness to unseen noise types is also one of the evaluation criteria. Thus, in this experiment, Aurora2 dataset [50] is incorporated as unseen test noises to further assess the robustness of seven models to various non-stationary noises. Tables 9 and Fig. 10 display the PESQ and STOI results of unprocessed and processed speech under different noisy-reverberant conditions by using different methods. Compared with results in Table 7 and Fig. 7, it is clear that when processing untrained noise types, the enhanced performance of seven models are all decreased, which implies that their robustness are declined more or less. However, the trends of their performance are similar to those of matching noises conditions but still have some difference. Specifically, the PESQ results in Table 9 state that MRCAE outperforms other models at low input SNRs (-5dB and 0dB), while C-LPS+IRM yields a superior performance at high input SNRs (5dB and 10dB). This may because in strong noise condition, the mapping- and masking- based methods remove the useful frequency domain information of clean speech during denoising and dereverberation, which in turn leads to speech distortion in time domain. Meanwhile, the average

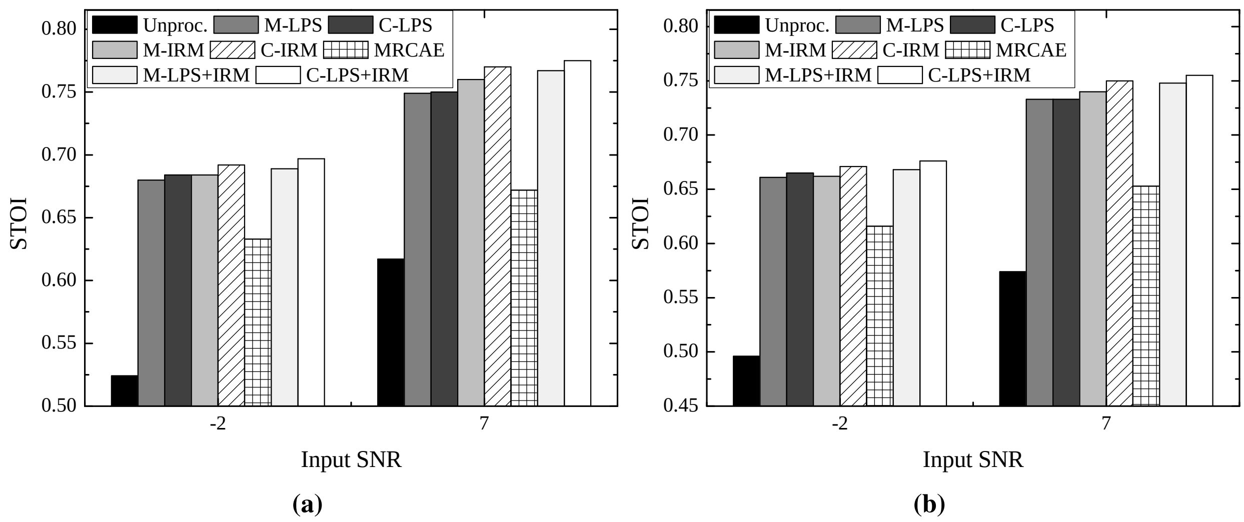



**Fig. 9.** Average STOI results in (a) Room 1 and (b) Room 2 at different untrained SNRs.

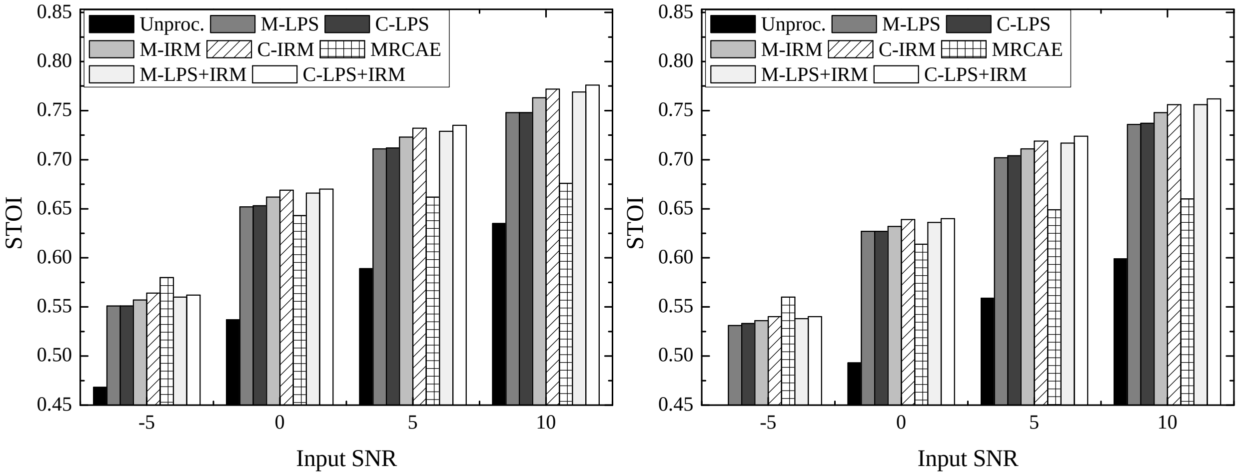
STOI results in Fig. 10 also reveal that only at low input SNR (-5dB), the performance of C-LPS+IRM is inferior to that of MRCAE. For other cases, C-LPS+IRM usually produces better STOI results than others. Thus, taken together, the proposed method has advantages in coping with various noise types and is suitable for different environments.

**Table 9** Average PESQ results with untrained noise types at different SNRs.

| | Room1 | | | | Room2 | | | |
|---|---|---|---|---|---|---|---|---|
| SNR(dB) | -5 | 0 | 5 | 10 | -5 | 0 | 5 | 10 |
| Unprocessed | 1.452 | 1.668 | 1.853 | 2.008 | 1.441 | 1.603 | 1.792 | 1.896 |
| M-LPS | 1.635 | 1.895 | 2.089 | 2.224 | 1.616 | 1.842 | 2.056 | 2.179 |
| C-LPS | 1.634 | 1.895 | 2.095 | 2.233 | 1.621 | 1.852 | 2.066 | 2.182 |
| M-IRM | 1.658 | 1.948 | 2.194 | 2.371 | 1.627 | 1.886 | 2.135 | 2.286 |
| C-IRM | 1.679 | 1.966 | 2.205 | 2.378 | 1.641 | 1.907 | 2.149 | 2.304 |
| MRCAE | **1.898** | **2.065** | 2.134 | 2.3198 | **1.850** | **1.957** | 2.074 | 2.100 |
| M-LPS+IRM | 1.664 | 1.954 | 2.194 | 2.371 | 1.639 | 1.895 | 2.148 | 2.298 |
| C-LPS+IRM | 1.663 | 1.964 | **2.214** | **2.386** | 1.643 | 1.904 | **2.163** | **2.310** |

### *3.3.5. Generalization to untrained dataset*

In order to comprehensively explore the generalization of the proposed method, an untrained dataset named Deep Noise Suppression (DNS) [51] is employed in this experiment to further evaluate the enhanced performance among seven models. The configurations of test condition are the same as Table 2 except that the noises are randomly selected from DNS wideband noise dataset. Besides, each of the clean speech is mixed with different noise types to fully validate the effectiveness of the proposed method. Table 10 and Fig. 11 show the average PESQ and STOI




**(a)** **(b)**

**Fig. 10.** Average STOI results in (a) Room 1 and (b) Room 2 with untrained noise types at different SNRs.

results in processing DNS dataset with different models. From Table 10, it can be seen that the PESQ results of dealing with DNS dataset are similar to those of untrained noise types in Table 9. The waveform-based method MRCAE has a better performance at low input SNRs. However, for the mapping- and masking- based methods, C-LPS+IRM still significantly outperforms other models under different acoustic conditions, confirming its generalization capability. Moreover, STOI results also exhibit encouraging performance of the proposed model, which is different from those of untrained noise types in Fig. 10. C-LPS+IRM stands out among all models, ranking the first. C-IRM has the second best, and MRCAE has the third rank only at low SNR input (-5dB). For other input SNRs (0dB, 5dB and 10dB), the performance of MRCAE are dramatically dropped, which indicates that the applicability of MRCAE is limited.

**Table 10** Average PESQ results with DNS dataset at different SNRs.

| | Room1 | | | | Room2 | | | |
|---|---|---|---|---|---|---|---|---|
| SNR(dB) | -5 | 0 | 5 | 10 | -5 | 0 | 5 | 10 |
| Unprocessed | 1.401 | 1.597 | 1.768 | 1.893 | 1.394 | 1.546 | 1.708 | 1.805 |
| M-LPS | 1.590 | 1.852 | 2.063 | 2.215 | 1.596 | 1.838 | 2.027 | 2.216 |
| C-LPS | 1.595 | 1.872 | 2.076 | 2.213 | 1.601 | 1.839 | 2.042 | 2.178 |
| M-IRM | 1.572 | 1.901 | 2.161 | 2.358 | 1.569 | 1.857 | 2.100 | 2.269 |
| C-IRM | 1.607 | 1.923 | 2.187 | 2.373 | 1.602 | 1.884 | 2.120 | 2.295 |
| MRCAE | **1.850** | **2.004** | 2.112 | 2.166 | **1.838** | **1.966** | 2.041 | 2.090 |
| M-LPS+IRM | 1.605 | 1.943 | 2.171 | 2.351 | 1.596 | 1.871 | 2.111 | 2.288 |
| C-LPS+IRM | 1.616 | 1.945 | **2.195** | **2.381** | 1.616 | 1.903 | **2.133** | **2.301** |

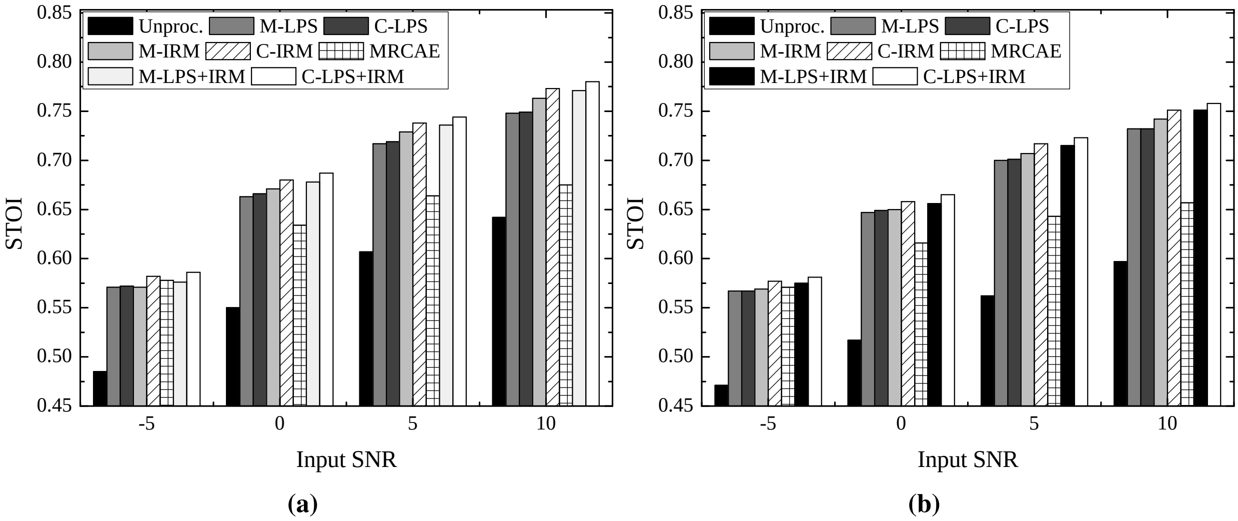



**Fig. 11.** Average STOI results in (a) Room 1 and (b) Room 2 with DNS dataset at different SNRs.

### *3.3.6. Generalization to untrained source-array distances*

In this experiment, two source-array distances (1.2m and 1.7m) that are not used in training are selected to further explore the generalization capability of the proposed model. All the other conditional configurations, including the noise types and input SNRs, are set as same as training stage. Table 11 shows the average STOI and PESQ results with different source-array distances among seven models. In general, no matter in short or long distance, C-LPS+IRM always exhibits substantial improvements, ranking the first; IRM and MS-LPS+IRM are both inferior to C-LPS+IRM, having the second best and the third rank. By observing the results in details, when the distance is 1.2m, the average PESQ increment of C-LPS+IRM is 0.014 compared to the second best C-IRM; However, when the distance is 1.7m, the increment of C-LPS+IRM is considerably increased to 0.027, which is almost twice compared to that of 1.2m distance. Meanwhile, STOI results shows that for short source-array distance, C-LPS+IRM has an average 0.148 improvements, which just slightly better than C-IRM with 0.146; But for long source-array distance, the STOI improvement of C-LPS+IRM is increased to 0.189, and yields 0.7 percent superior than that of C-IRM. In addition, it can be found that for STOI metric, the advantage of proposed C-LPS+IRM are more obvious in Room 2 (strong noisy-reverberation condition). The improvements achieve to 0.148 and 0.206 which comparable or even better than 0.148 and 0.171 in Room 1.

From above analyses, it can be concluded that the proposed C-LPS+IRM models significantly improves the dereverberation and denoising performance over unprocessed signals in untrained acoustic conditions, and this indicates that it has great prospect for speech enhancement under severe environments.

## 4. Conclusion

In order to address both the non-Gaussian noise and reverberation problems, we propose a multi-objective based multi-channel speech enhancement network by using correntropy as the

**Table 11** Average STOI and PESQ results with different source-array distances.

| | PESQ | | | | STOI | | | |
|---|---|---|---|---|---|---|---|---|
| | Room1 | | Room2 | | Room1 | | Room2 | |
| Distance(m) | 1.2 | 1.7 | 1.2 | 1.7 | 1.2 | 1.7 | 1.2 | 1.7 |
| Unprocessed | 1.755 | 1.744 | 1.719 | 1.655 | 0.586 | 0.546 | 0.590 | 0.480 |
| M-LPS | 2.133 | 2.121 | 2.123 | 2.058 | 0.711 | 0.702 | 0.710 | 0.676 |
| C-LPS | 2.137 | 2.133 | 2.136 | 2.063 | 0.713 | 0.704 | 0.712 | 0.679 |
| M-IRM | 2.216 | 2.185 | 2.215 | 2.087 | 0.722 | 0.701 | 0.728 | 0.668 |
| C-IRM | 2.236 | 2.200 | 2.227 | 2.109 | 0.731 | 0.711 | 0.736 | 0.679 |
| MRCAE | 2.075 | 2.010 | 2.117 | 1.934 | 0.667 | 0.618 | 0.643 | 0.597 |
| M-LPS+IRM | 2.229 | 2.206 | 2.221 | 2.115 | 0.727 | 0.710 | 0.729 | 0.679 |
| C-LPS+IRM | **2.249** | **2.229** | **2.241** | **2.133** | **0.734** | **0.717** | **0.738** | **0.686** |

loss function. First, the log-power spectra (LPS) of multiple channels noisy speech are employed as the input of BiLSTM network. Then, multiple correntropy-based loss functions are used to simultaneously estimate the intermediate LPS and log-ideal ratio mask (LIRM) of each channel. Subsequently, two fusion layers are adopted to separately fuse the intermediate LPS and LIRM into two single-channel features, which are then integrated into a single-channel LPS. Ultimately, a deep neural network is brought in to further approximate the nonlinear mapping from the final fused LPS to the clean speech LPS. Moreover, during training, the correntropy based loss function is employed due to its insensitivity to the outliers (or impulsive noise) for the sake of improving the network performance. Experimental evaluations show that the proposed method has the superiority in suppressing non-Gaussian noise and exhibits good generalization capability to different noises, SNRs, datasets and source-array distances.


## Acknowledgments

This work was supported by National Natural Science Foundation of China (Nos. 61771091, 61871066), National High Technology Research and Development Program (863 Program) of China (No. 2015AA016306), Natural Science Foundation of Liaoning Province of China (No. 20170540159), and Fundamental Research Funds for the Central Universities of China (Nos. DUT17LAB04).


## Data Availability Statement

The datasets generated during the current study are available from the corresponding author on reasonable request.

## Conflict of Interest Statement

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

## References


[1] J. Li, L. Deng, R. Haeb-Umbach, Y. Gong, Robust Automatic Speech Recognition: A bridge to practical applications. Academic Press, 2015.

[2] J. Benesty, S. Makino, J. Chen, Speech Enhancement. Springer, 2005.

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

[4] S. Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27(2), 113-120 (1979).

[5] J. S. Lim, A. V. Oppenheim, Enhancement and bandwidth compression of noisy speech. Proc. of IEEE 67(12), 1586-1604 (1979).

[6] Y. Ephraim, H. L. Van Trees, A signal subspace approach for speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 3(4), 251-266 (1995).

[7] I. Cohen, S. Gannot, Springer Handbook of Speech Processing. Berlin, Germany: Springer-Verlag, 2008.

[8] T. Yoshioka, T. Nakatani, Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. IEEE Trans. Acoust Speech Signal Process. 20(10), 2707-2720 (2012).

[9] S. T. Neely, J. B. Allen, Invertibility of a room impulse response. J. Acoust. Soc. Amer. 66, 165-169 (1979).

[10] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, T. Zhang, Learning spectral mapping for speech dereverberation and denoising. IEEE/ACM Trans. Audio Speech Lang. Process. 23(6), 982-992 (2015).

[11] D. S. Williamson, D. L. Wang, Time-frequency masking in the complex domain for speech dereverberation and denoising. IEEE/ACM Trans. Audio Speech Lang. Process. 25(7), 1492-1501 (2017).

[12] Y. Zhao, Z. Wang, D. Wang, Two-stage deep learning for noisy-reverberant speech Enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 27(1), 53-62 (2019).

[13] R. Li, X. Sun, T. Li, F. Zhao, A multi-objective learning speech enhancement algorithm based on IRM post-processing with joint estimation of SCNN and TCNN. Digit. Signal Process. 101 (2020).

[14] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, A consolidated perspective on multi-microphone speech enhancement and source separation. IEEE/ACM Trans. Audio Speech Lang. Process. 25(4), 692-730 (2017).

[15] N. Yousefian, P. C. Loizou, A dual-microphone speech enhancement algorithm based on the coherence function. IEEE/ACM Trans. Audio Speech Lang. Process. 20(2), 599-609 (2012) .

[16] T. Shan, T. Kailath, Adaptive beamforming for coherent signals and interference. IEEE Trans. Acoust. Speech Signal Process. 33(3), 527-536 (1985).

[17] I. Tashev, A. Acero, Microphone array post-processor using instantaneous direction of arrival, in *International Workshop on Acoustic, Echo and Noise Control (IWAENC)*, (2006).

[18] Y. Wang, A. Narayanan, D. Wang, On training targets for supervised speech separation. IEEE Trans. Acoust. Speech Signal Process. 22(12), 1849-1858 (2014).

[19] Y. Jiang, D. Wang, R. Liu, Z. Feng, Binaural classification for reverberant speech segregation using deep neural networks. IEEE Trans. Acoust. Speech Signal Process. 22(12), 2112-2121 (2014).

[20] X. Zhang and D. Wang, Deep learning based binaural speech separation in reverberant environments. IEEE Trans. Acoust. Speech Signal Process. 25(5), 1075-1084 (2017).

[21] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, T. Nakatani, Exploring multi-channel features for denoising-autoencoder-based speech enhancement, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 116-120 (2015).

[22] X. Sun, R. Xia, J. Li, Y. Yan, A deep learning based binaural speech enhancement approach with spatial cues preservation, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5766-5770 (2019).

[23] J. Heymann, L. Drude, R. Haeb-Umbach, Neural network based spectral mask estimation for acoustic beamforming, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 196-200 (2016).

[24] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, H. Li, On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3246-3250 (2017).

[25] S. Chakrabarty, D. Wang, E. A. P. Habets, Time-frequency masking based online speech enhancement with

multi-channel data using convolutional neural networks, in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 476-480 (2018).
[26] S. Chakrabarty, E. A. P. Habets, Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks. IEEE J. Sel. Topics Signal Process. 13(4), 787-799 (2019).
[27] X. Cui, Z. Chen, F. Yin, Multi-objective based multi-channel speech enhancement with BiLSTM network. Appl. Acoust. 177, (2021).
[28] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, T. Nakatani, Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 531-535 (2018).
[29] J. Qi, H. Hu, Y. Wang, C. H. Yang, S. Marco Siniscalchi, C. Lee, Tensor-to-vector regression for multi-channel speech enhancement based on tensor-train network, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7504-7508 (2020).
[30] I. Santamaria, P. P. Pokharel, J. C. Principe, Generalized correlation function: Definition, properties, and application to blind equalization. IEEE Trans. Signal Process. 54(6), 2187-2197 (2006).
[31] W. Liu, P. P. Pokharel, J. C. Principe, Correntropy: Properties and applications in non-Gaussian signal processing. IEEE Trans. Signal Process. 55(11), 5286-5298 (2007).
[32] P. P. Pokharel, W Liu, J. C. Principe, A low complexity robust detector in impulsive noise. Signal Process. 89(10), 1902C1909 (2009).
[33] A. Singh, J. C. Principe, A loss function for classification based on a robust similarity metric, in *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-6 (2010).
[34] A. Singh, R. Pokharel, J. C. Principe, The c-loss function for pattern classification. Pattern Recognit. 47(1), 441-453 (2014).
[35] Y. Qi, Y. Wang, X. Zheng, Z. Wu, Robust feature learning by stacked autoencoder with maximum correntropy criterion, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6716-6720 (2014).
[36] L. Chen, H. Qu, J. Zhao, B. Chen, J. C. Principe, Efficient and robust deep learning with correntropy-induced loss function. Neural Computing Appl. 27(4), 1019-1031 (2016).
[37] D. Wang, J. Chen, Supervised speech separation based on deep Learning: An overview. IEEE Trans. Acoust Speech Signal Process. 26(10), 1702-1726 (2018).
[38] K. Greff, R. K. Srivastava, J. Koutnłk, B. R. Steunebrink, J. Schmidhuber, LSTM: A search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. 28(10), 2222-2232 (2017).
[39] S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. 9(8), 1735-1780 (1997).
[40] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 18(5-6), 602-610 (2005).
[41] L. Sun, J. Du, L. Dai, C. Lee, Multiple-target deep learning for LSTM-RNN based speech enhancement, in *Hands-Free Speech Communications and Microphone Arrays*, (HSCMA), pp. 136C140 (2017).
[42] J. S. Garofolo, L. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, Darpa TIMIT acoustic-phonetic continuous speech corpus. 1993. [Online]. Available: https://github.com/philipperemy/timit
[43] ] E. A. P. Habets, Room impulse response (RIR) generator. 2016. [Online]. Available: https://github.com/ehabets/RIR-Generator
[44] A. Varga, H. 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 Commun. 12(3), 247-251 (1993).
[45] F. Chollet *et al.*, Keras. 2015. [Online]. Available: https://github.com/fchollet/keras
[46] E. M. Grais, D. Ward, M. D. Plumbley, Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders, in *European Signal Processing Conference (EUSIPCO)*, pp. 1577-1581 (2018).
[47] ITU-T, Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Technical Report, 2001.
[48] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, A short-time objective intelligibility measure for time-frequency weighted noisy speech, in *IEEE International Conference on Acoustics, Speech and Signal Processing*

*(ICASSP)*, pp. 4214-4217 (2010).

[49] J. Jensen, C. H. Taal, An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE Trans. Acoust Speech Signal Process. 24(11), 2009-2022 (2016).

[50] H. G. Hirsch, D. Pearce, The AURORA experimental framework for the preformance evaluations of speech recognition systems under noisy conditions, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 181-188 (2000).

[51] C. K. A. Reddy *et al.*, The Interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework, 2020.

Dear Editor,

We have submitted our revised manuscript entitled "***Correntropy based Multi-objective Multi-channel Speech Enhancement***" by Xingyue Cui *et al.*, with the paper ID: CSSP-D-21-00332. The editor's and reviewer's comments are very valuable and helpful for improving our manuscript.

The following pages are the point-by-point responses to the reviewers' comments, and the corresponding modifications have been made and highlighted with red mark in the revised manuscript. We hope that this manuscript can be considered for publication in *Circuits, Systems, and Signal Processing.*

If have any problem, please do not hesitate to contact us. Thanks.

On behalf of all authors,

Prof. Zhe Chen

Tel: +86-411-84707049
E-mail: zhechen@dlut.edu.cn
School of Information and Communication Engineering,
Dalian University of Technology,
Dalian 116023, P.R. China

**Correntropy based Multi-objective Multi-channel Speech Enhancement (ID: CSSP-D-21-00332)**

Xingyue Cui, Zhe Chen, Fuliang Yin, Xianfa Xu

## The Point-by-point Responses to the Reviewer's Comments

We would like to express our appreciation to the reviewers for providing us with valuable comments for improving this manuscript. In the following, we present the point-by-point replies to the reviewers' comments.

## Reviewer1:

**Comments No.1:** *However, the authors show the experimental results, which evident the claim, but an analytical claim for the same is missing. Better to provide a subsection, and explain why correntropy loss performs better with non Gaussian noise (appropriate reference for the same is also missing). The computation of IRM features is also not so clear, expected to provide a block diagram for the same.*

**Answer:**

Thanks for the reviewer's valuable advice. Generally, non-Gaussian noises, especially impulse noise, are prone to large outliers during training process. Correntropy loss has the capability to suppress the outliers because of the nonlinear kernel function, such as Gaussian kernel function. According to the characteristic of Gaussian function, the lager the outliers, the better the suppression. In contrast, MSE loss would amplify the outliers during training, and is not optimal in the case of asymmetric error distribution, non-zero center and outliers. In other words, nonlinear kernel function is the key factor that makes correntropy robust to outliers, and it is the reason why correntropy loss performs better with non-Gaussian noise.

According to the reviewer's comment, we have added the above reason and reference in the revised manuscript. Please refer to the fourth paragraph of Subsection 2.5. Besides, a block diagram of the computation of IRM feature has been also added in the revised manuscript. Please refer to Fig. 1 of Subsection 2.3.

**Comments No.2:** *The motivation for the work is well written, If the authors can summarize the motivation through a table or Block diagram, it makes the article more readable.*

**Answer:**

According to the reviewer's advice, we have summarized the motivation through a table. Please refer to the Table 1 of Section 1 in the revised manuscript.

**Comments No.3:** *"First, the log-power spectra (LPS) of multichannel noisy speech* *are used as the input of the bi-directional long short-term memory (BiLSTM) network to predict the ideal ratio mask (IRM) and LPS of clean speech in each channel.", In the abstract, the author may re-frame the sentence, it is not so clear.*

**Answer:**

According to the reviewer's advice, we have modified the sentence "*First, the log-power spectra (LPS) of multichannel noisy speech are used as the input of the bi-directional long short-term memory (BiLSTM) network to predict the ideal ratio mask (IRM) and LPS of clean speech in each channel.*" to "First, the log-power spectra (LPS) of multichannel noisy speech are feed to the bidirectional long short-term memory (BiLSTM) network with the aim of predicting the intermediate ideal ratio mask (IRM) and LPS of clean speech in each channel." with red font in the revised manuscript.

**Comments No.4:** *Page no. 4, line 26, the abbreviation of SFTF is missing.*

**Answer:**

Sorry for our carelessness. We have added the abbreviation of SFTF with red font in the revised manuscript.

**Comments No.5:** *In section 2.3, it will be better to explain why clean speech signal magnitude is less than the magnitude of noisy signal. (Ideally)*

**Answer:**

Thanks for the reviewer's comment. Section 2.3 describes ideal ratio mask which is defined as $IRM(n,k)=\frac{|S(n,k)|}{|X(n,k)|}$ in Eq.(4), where S($n$,$k$) and X($n$,$k$) denote the clean- and noisy- spectral magnitudes within a T-F unit, respectively. In fact, S($n$,$k$) is not always less than X($n$,$k$), and the values of IRM may be greater than one. However, a larger value may lead to instability of backpropagation during training, thus, the maximum value of IRM is set to 1 in our method, i.e. S($n$,$k$) has to be no greater than X($n$,$k$). We have explained the reason in the revised manuscript, please refer to the sentences below Eq.(4) in Section 2.3.

## Reviewer2:

**Comments No.1:** *The main contribution of the manuscript is exploring Correntropy as the loss function for non-Gaussian distribution. The selection of the kernel size is very crucial to get improve performance compared to other methods. In the manuscript the authors have chosen kernel size based on PESQ and different SNR.*

*However, I have a concern with the selection of the kernel function i.e., Gaussian kernel in the manuscript. Although Gaussian kernel is desirable, but the correntropy induced metric varies with the data and kernel size. In the literature, methods are presented to avoid choosing kernel size and match the sample distribution. Therefore, a proper justification for this concern is expected from the authors.*

**Answer:**

The reviewer raises an interesting concern. Reference [A1] ([30] in revised manuscript) first proposed the concept of correntropy which is based on the kernel method and information theoretic learning theory. It is pointed that the basic idea of kernel method is to transform the data $x_i$ from the input space to a high dimensional feature space of vectors $\phi(x_i)$, where the inner products can be computed using a positive definite kernel function satisfying Mercer's conditions: $\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. Thus, without loss of generality, the translation-invariant Gaussian kernel, which is the most widely used Mercer kernel, was employed when computing correntropy [A1, A2] ([30, 32] in revised manuscript). As we know, the kernel function of correntropy is almost Gaussian kernel. Although References [A3, A4] proposed mixture correntropy and multi-kernel correntropy for robust learning, the essence of these methods is to combine several Gaussian functions to construct the kernel function. Therefore, based on common sense and above experiences, we select Gaussian kernel as the kernel function.

[A1] *I. Santamaria, P. P. Pokharel, J. C. Principe, Generalized correlation function: Definition, properties, and application to blind equalization. IEEE Trans. Signal Process. 54(6), 2187-2197 (2006).*

[A2] *P. P. Pokharel, W Liu, J. C. Principe, A low complexity robust detector in impulsive noise. Signal Process. 89(10), 1902–1909 (2009).*

[A3] *B. Chen, X. Wang, N. Lu, S. Wang, J. Cao, J. Qin, Mixture correntropy for robust learning, Pattern Recognit., 79, 318–327 (2018).*

[A4] *B. Chen, X. Wang, Z. Yuan, P. Ren, J. Qin, Multi-kernel correntropy for robust learning, 2019.[Online]. Availabvle: http://arxiv.org/abs/1905.10115v1.*

**Comments No.2:** *In the manuscript, the symbol of 'sigma' is used to refer sigmoid function and Gaussian kernel. If possible try to differentiate between the two.*

**Answer:**

Sorry for our carelessness. We have modified the $\sigma$ symbol that refers to the sigmoid function to $\kappa$. Please refer to Eqs. (5), (6) and (8) with red font in the revised manuscript.

**Comments No.3:** *In addition to PESQ and STOI, the ESTOI value could also be computed to reveal the spectral deviations and improvement after mapping the noisy* *speech to the approximate clean speech.*

**Answer:**

According to the reviewer's advice, the ESTOI metric has been added when evaluating the performance of speech enhancement. Please refer to Fig. 6 and the second paragraph of Subsection 3.3.2 1) in the revised manuscript.

**Comments No.4:** *Overall the manuscript need to go through proofreading to improve the grammatical errors, typos, and English. Typo in Fig. 2 must be corrected.*

**Answer:**

Thanks for the reviewer's advice. We have checked the grammatical errors, typos and English throughout the manuscript carefully, and further improved English quality of our manuscript. Typos in Fig. 2 (Fig. 3 in revised manuscript) have been corrected, please refer to page 9.

**Comments No.5:** *Use of abbreviations are not consistent. If you are using abbreviations, define it the first time you are using and do not expand it next time you use it in the manuscript.*

**Answer:**

Thanks for the reviewer's advice. We have carefully checked the use of abbreviations throughout the manuscript.

**Comments No.6:** *Some abbreviations are mis-handled, the authors must avoid such mistake, for e.g., SFTF for STFT and MMC for MCC.*

**Answer:**

Sorry for our carelessness. We have carefully corrected the spelling mistakes of the abbreviations throughout the manuscript. The "SFTF" and "MMC" have been modified to "STFT" and "MCC" with red font in the revised manuscript.

**Comments No.7:** *In subsection 2.3 please provide relevant references for which you are referring to the recent studies.*

**Answer:**

According to the reviewer's advice, we have cited the relevant references in the revised manuscript. Please refer to the references [18, 37] in subsection 2.3.

**Reviewer3:**

**Comments No.1:** *The authors should clearly state how their current proposal is different from the previous work [27], which is also based on multi-objective based multi-channel speech enhancement. A point-wise contribution of the current work is required apart from showing the differences from the previous work.*

**Answer:**
Thanks for the reviewer's suggestion. Indeed, the current proposal is an extension of the previous work [27], but these two methods have three main differences. First, the loss function is different. This work proposed a model based on correntropy loss function, while the model in Ref. [27] used MSE as loss function. Second, the datasets used for training and testing are different. In this work, although clean and noise speech are the same as Ref. [27], the way of constructing noisy speech is totally different. In addition, the acoustic conditions have also extended to more severe conditions. As shown in Table I and Table II, the RT60 are set to {0.3s, 0.4s, 0.6s, 0.8s, 0.9s} during training, and {0.45s, 0.7s} during testing. But in Ref. [27], the maximum of RT60 was only set to 0.6s and 0.57s during training and testing, respectively. Third, the model structure is different. This work applied a BiLSTM to learn multi-channel multi-objective function simultaneously, while Ref. [27] employed a BiLSTM with shared weights, which processes the input of each channel separately. Besides, after the LPS fusion layer, a correntropy based loss function is also added to further boost the learning capability of the proposed model.

Therefore, the main contributions of this manuscript are summarized as follows: 1) A correntropy based multi-objective multi-channel speech enhancement model is proposed, i.e. correntropy is employed to construct a multi-objective loss function to optimize the model; 2) The model structure is designed more reasonable, and displays a superior enhanced performance in untrained acoustic conditions; 3) More severe acoustic conditions have been considered during training to improve the generalization of the proposed model.

According to the reviewer's comment, we have added the above-mentioned point-wise contribution of the current work in the revised manuscript. Please refer to the penultimate paragraph in Section 1.

**Comments No.2:** *The authors do not show any illustration of noisy speech considered and the enhanced speech in terms of waveforms or spectrograms, which would have been useful for the readers. Additionally, if they could release some of the noisy speech examples and corresponding enhanced samples for the readers.*

**Answer:**
According to the reviewer's advice, the waveforms and spectrograms, and the corresponding description of analyses have been added to further demonstrate the

effectiveness of the proposed method. Please refer to Fig. 8 and the last paragraph of Subsection 3.3.2 2) in the revised manuscript. Besides, the noisy speech examples and corresponding enhanced speech samples are released in “http://faculty.dlut.edu.cn/chenzhe/zh_CN/jxzy/745520/content/3834.htm#jxzy”.

**Comments No.3:** *The authors have used TIMIT database for studies with additional noises. It would have been nice if the methods could have been validated using DNS challenge data, which is recently used for speech enhancement research.*

**Answer:**

According to the reviewer’s suggestion, the wide band track of DNS challenge data have been used to validated the generalization of the proposed method. Please refer to the Subsection 3.3.5 in the revised manuscript.

**Comments No.4:** *The authors do not compare their methods with some of the latest state-of-the-art methods such as MetricGAN (https://github.com/speechbrain/speechbrain). It is always good to compare the proposed method with some latest methods.*

**Answer:**

Thanks for the reviewer’s advice. MetricGAN is designed to solve the discriminator-evaluation mismatch (DEM) problem in GAN, and presents superior performance in speech enhancement task. However, it is a single-channel based method, and we think it is unfair to compare it with our multi-channel based method. Therefore, a latest multi-channel method named multi-resolution convolutional auto-encoders (MRCAE) [A1] ([46] in revised manuscript) is employed for comparison, and the results have been added in all the experiments. Please refer to the Tables 6-11 and Figs. 5-11 and the corresponding descriptions of Subsections 3.2, 3.3 in the revised manuscript.

[A1] *E. M. Grais, D. Ward, M. D. Plumbley, Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders, in: European Signal Processing Conference (EUSIPCO), pp. 1577-1581 (2018).*

**Comments No.5:** *In Section 3.3.1, it is mentioned "From Table 3, when the value of \sigma 1.0, 1.3, 1.4 and 1.5, the performance of the model tends to be the best." However, the reviewer finds that when \sigma =1.0, it is not the best. The PESQ value at 1.8 is better than 1.0. Thus, it is a bit strange to consider \sigma=1.0 for the next study shown in Table 4 and then to report it as the best case.*

**Answer:**

Sorry for our unclear description. Table 3 shows the average PESQ results with different kernel sizes $\sigma$. When the value of $\sigma$ is 1.0, 1.3, 1.4 and 1.5, the corresponding PESQ is 2.219, 2.218, 2.218, and 2.218, respectively. However, when the value of $\sigma$ is 1.8, the PESQ is only 2.217. Thus, we mentioned that "From Table 3, when the value of $\sigma$ 1.0, 1.3, 1.4 and 1.5, the performance of the model tends to be the best.", and selected 1.0 as the value of $\sigma$ in the following experiments.

**Comments No.6:** *In Section 3.3.2, machinegun noise is used for study. Instead of machinegun noise, some commonly occurring noise like vehicle noise would be nice to observe.*

**Answer:**
In our manuscript, three noise datasets (NOISEX-92, Aurora2 and DNS) are employed to evaluate the enhanced performance of the proposed method. They contain the commonly occurring noise, such as vehicle noise. From the results of overall performance in Table 7 and Fig. 7, untrained noise types in Table 9 and Fig. 10 and DNS dataset in Table 10 and Fig. 11, it can be observed that the proposed method has good robustness to commonly occurring noises. Besides, machinegun noise is separately used only to study the advantage of the proposed method under impulse noise. Therefore, considering the page limitation, we did not conduct another vehicle noise experiment in the revised manuscript.

**Comments No.7:** *The reviewer guessed in page 13, overall performance evaluation should be a new subsection.*

**Answer:**
Yes, "overall performance evaluation" is a new subsection. To make it clearer, we have added a label for this subsection. Please refer to Subsection 3.2.2 2) in the revised manuscript.

**Comments No.8:** *In Table 8, why did C-IRM perform the best for low SNR input?*

**Answer:**
Table 8 shows the average PESQ results with untrained noise types at different SNRs. It can be seen that the waveform-based method MRCAE has the best performance. However, for the mapping- and masking- based methods, C-IRM performs the best for low SNR input, and the reasons are summarized as follows. First, when the input SNR is low, the clean speech are almost submerged in noises, which makes the predicted errors of multi-objective based methods lager than those of single-objective based methods. Second, compared with LPS feature, IRM estimates the probability of speech at each frequency, and has the capability to retain more useful information of clean speech and reduce speech distortion in time domain. Third, the loss function based on correntropy is more insensitivity to outliers (impulse noise) than those based on MSE. Based on above reasons, C-IRM performs the best among the mapping- and masking- based methods at low SNR input.

**Comments No.9:** *There are several typos in the paper. Authors should read carefully and revise. Some common typos are:*

*(i) page 2, level differences (ILD)) => level differences (ILD)*

*(ii) page 3, we extend our previous work [27] to proposed a ... => we extend our previous work [27] to propose a ....*

*(iii) page 5, Generally, its value is range from ... => Generally, its value ranges from ...*

**Answer:**

Thanks for the reviewer's kind comment. We have carefully corrected the careless typos throughout this manuscript. The "*level differences (ILD))*" has been modified to "*level differences (ILD)*" , "*we extend our previous work [27] to proposed a ...*" has been modified to "*we extend our previous work [27] to propose a*…" and "*Generally, its value is range from* …" has been modified to "*Generally, its value ranges from ...*" with red font in the revised manuscript. Besides, other several typos have also been checked carefully and corrected totally in the revised manuscript.