Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 4618317, 14 pages
https://doi.org/10.1155/2020/4618317

*Research Article*

# Deep Learning-Based Amplitude Fusion for Speech Dereverberation

**Chunlei Liu,[1,2] Longbiao Wang,[2] and Jianwu Dang [2,3]**

[1]*School of Computer and Information, Dezhou University, Dezhou 253023, China*
[2]*Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China*
[3]*Japan Advanced Institute of Science and Technology, Ishikawa 9231292, Japan*

Correspondence should be addressed to Jianwu Dang; jdang@jaist.ac.jp

Received 21 February 2020; Revised 22 May 2020; Accepted 20 June 2020; Published 14 July 2020

Academic Editor: Florentino Borondo



Mapping and masking are two important speech enhancement methods based on deep learning that aim to recover the original clean speech from corrupted speech. In practice, too large recovery errors severely restrict the improvement in speech quality. In our preliminary experiment, we demonstrated that mapping and masking methods had different conversion mechanisms and thus assumed that their recovery errors are highly likely to be complementary. Also, the complementarity was validated accordingly. Based on the principle of error minimization, we propose the fusion between mapping and masking for speech dereverberation. Specifically, we take the weighted mean of the amplitudes recovered by the two methods as the estimated amplitude of the fusion method. Experiments verify that the recovery error of the fusion method is further controlled. Compared with the existing geometric mean method, the weighted mean method we proposed has achieved better results. Speech dereverberation experiments manifest that the weighted mean method improves PESQ and SNR by 5.8% and 25.0%, respectively, compared with the traditional masking method.

## 1. Introduction

In the real-world speech environment, original clean speech is often corrupted by reverberation, which seriously damages the speech quality and reduces the performance of automatic speaker recognition [1] and automatic speech recognition (ASR) [2, 3]. When the reverberation is too heavy, human hearing also suffers severe interference [4]. Speech enhancement refers to the processing with corrupted speech to obtain underlying clean speech, thereby improving the speech quality. It mainly includes speech dereverberation and speech denoising. This paper focuses on speech dereverberation. In the early stage, unsupervised speech enhancement methods were usually used to improve the corrupted speech. The traditional speech dereverberation method is mainly the weighted prediction error (WPE) method [5, 6], but the improvement of speech in this method is rather limited. With the rapid development of deep learning in recent years, supervised speech enhancement methods were emerged. Two types of these methods are especially important. One is called mapping [7–10], which uses a deep neural network (DNN) to directly predict clean speech feature from corrupted speech feature (MAP). The feature here usually refers to the logarithmic power spectrum (LPS), which conforms to human auditory rules and is beneficial to the learning of DNN. The other is called masking [11, 12], which predicts an intermediate state from the corrupted speech feature, and the clean speech is obtained from the known corrupted speech and the predicted intermediate state. Because DNN has a strong learning ability, mapping and masking methods can effectively reduce the reverberation components in the reverberant speech. The masking methods include the ideal binary mask (IBM) [13, 14], the ideal ratio mask (IRM) [15, 16], the ideal amplitude mask (IAM) [11, 17] (known as FFT-MASK in [11]), the phase-sensitive mask (PSM) [18], and the complex

ratio mask (cIRM) [19–22]. IRM takes advantage of the incoherence between clean speech and additive noise to approximate the power of the corrupted speech to the sum of the power of the original clean speech and additive noise, thus limiting the IRM training target to [0; 1]. This facilitates the training of DNN and makes IRM a great success in reducing additive noise.

The authors of [23] pointed out that MAP and IRM were complementary in speech denoising under different signal-to-noise ratio (SNR); that is, MAP is better than IRM when SNR is low, and MAP is below IRM when SNR is high. Similar complementarity study has also been reported in [24–28]. Based on this complementarity, the authors of [23] proposed a geometric mean method fusing the amplitudes of MAP and IRM to improve the speech denoising effect. Although this fusion method is exceedingly effective for speech denoising, it is not very favorable for speech dereverberation. Consequently, further effort is required to understand and better take advantage of the fusion method.

In this paper, we analyze the principle and mechanism of mapping and masking methods and assume that different conversion mechanisms will lead to the complementarity of speech enhancement effects. Based on this, we conclude that the complementarity between the mapping and masking methods is widespread. It is not limited to the MAP and IRM methods, and there is also complementarity between MAP and other methods belonging to masking. Since the IRM method has great limitations in speech dereverberation, we propose the fusion between IAM and MAP for speech dereverberation. In order to further explore the fusion mode that minimizes the recovered amplitude error, we propose the arithmetic mean and the weighted mean.

Clean speech will become into different corrupted speeches when interfered by different speech scenarios. Accordingly, the correspondence between corrupted speech and clean speech is actually "many-to-one." Therefore, MAP is one kind of "many-to-one" conversion. While the training target of masking is typically a time-frequency (T-F) amplitude ratio of clean speech to corrupted speech. Hence, it is easier to infer that masking is a "one-to-one" conversion. The recovery errors produced by different conversion mechanisms are more likely to be complementary. Specifically, one conversion mechanism can overestimate the clean speech amplitude, while another is likely to underestimate it, and we found this phenomenon in the preliminary experiments. Based on the principle of error minimization, the mean of the amplitudes recovered by the mapping and masking methods will further improve the recovery accuracy. However, the IRM method is not satisfactory in terms of speech dereverberation because convolutional reverberation and additive noise have different mechanisms on destroying clean speech. Early reverberation has a strong correlation with clean speech, not as assumed in IRM. The training target of IAM is the T-F amplitude ratio of clean speech to corrupted speech. It has no limitation on the formation mechanism of corrupted speech and has great potential in both speech denoising and speech dereverberation. Therefore, the fusion of IAM and MAP will be more conducive to speech dereverberation.

In summary, this paper proposes that the different conversion mechanisms of mapping and masking methods lead to the complementarity, which in turn minimizes the recovery errors of the fusion method, and the scope of fusion is extended from the fusion of MAP and IRM to the fusion of mapping and masking methods. For speech dereverberation, we propose the fusion of MAP and IAM, and a new effective fusion mode is also proposed: the weighted mean. The speech dereverberation experiments suggest that the proposed methods exceed the existing related methods. The other three contributions of this paper are as follows: (1) This paper proposed the novel target DCC, namely, the difference between constructed and masked outputs. (2) We propose to use the "standard deviation" to measure the error of the predicted amplitude and analyze that the "standard deviation" is more reasonable than the traditional "ratio." (3) We propose to use DNN to predict the weight coefficients in the weighted mean, and the effect is very good. The training targets (weight coefficients) are different from the traditional training target directly labeled though the corpus (such as mapping and masking) but are obtained from the outputs of the DNN.

The remaining parts of this paper are organized as follows. The mapping and masking methods are introduced in Section 2. In Section 3, the amplitude fusion method is described. In Section 4, analysis of conversion mechanism and error minimization is provided. In Section 5, the speech dereverberation experiments and discussions are carried out, and conclusions are provided in Section 6.

## 2. Mapping and Masking Methods

When speech is corrupted by both reverberation and additive noise simultaneously, its mathematical expression in the time domain can be

$$y(t) = o(t) * r(t) + n(t). \tag{1}$$

Here, $y$, $o$, $r$, and $n$ refer to corrupted speech, original clean speech, room impulse response (RIR), and additive noise, respectively. t indexes the time, and $*$ represents the convolution. The short-time Fourier transform (STFT) of (1) is

$$Y(m, f) = O(m, f) \cdot R(m, f) + N(m, f). \tag{2}$$

Here, $Y$, $O$, $R$, and $N$ refer to the STFT of corrupted speech, original clean speech, RIR, and additive noise, respectively, while $m$ and $f$ index the time frame and frequency bin, respectively. "$\cdot$" represents the multiplication operation. The first term on the right side of (1) or (2) represents reverberant speech, while the second term represents additive noise. It can be seen that the influence mechanisms of reverberation and additive noise are very different. The task of speech enhancement is to estimate the underlying clean speech $o(t)$ from the corrupted speech $y(t)$.

Mapping has one classic form: predicting the LPS of clean speech directly from the LPS of corrupted speech using a DNN (MAP). The regularly used LPS is numerically twice as large as the log magnitude spectrum (LMS) used in this

paper. In our experiment, the speech quality obtained by LMS is not worse than that by LPS. The value of LMS can theoretically be in $(-\infty, +\infty)$, but the actual value is mostly in (−10, 4) within our preliminary experimental dataset. Accordingly, the output activation function of MAP is linear. The amplitude of the target clean speech is obtained by the exponential expansion of the estimated LMS.

The masking method has a variety of forms, with IRM being the most widely used method to reduce additive noise, and IAM has great potential in reducing both additive noise and convolutional reverberation. The training target of IAM is the T-F amplitude ratio of clean speech to corrupted speech [11], and the expression is as follows:

$$M^{\text{iam}}(m, f) = \frac{|O(m, f)|}{|Y(m, f)|}. \tag{3}$$

Here, $M^{\text{iam}}$ denotes the training target of IAM. $|\cdot|$ represents the modulus of the parameter. $M^{\text{iam}}$ is a typical training target of the masking method, through which a few variants can be obtained. The theoretical value range of $M^{\text{iam}}$ is $[0, +\infty)$; thus the output activation function of IAM is linear. Too large values of $M^{\text{iam}}$ are often clipped to facilitate the training of DNN, just as [11] advises that all values greater than 10 are set to 10. The amplitude of the target clean speech is obtained by multiplying the estimated value of $M^{\text{iam}}$ with the amplitude of the corrupted speech.

Historically, the IRM is extended from the ideal binary mask (IBM) [29]. But in this paper, we intentionally regard IRM as a variant of IAM. In the IRM method, we actually have an approximation as a prerequisite; that is, the power of noisy speech is set as the sum of the power of clean speech and noise. The T-F expression of the noisy speech corrupted by additive noise is

$$Y(m, f) = O(m, f) + N(m, f). \tag{4}$$

And the approximate expression of the noisy speech amplitude is

$$\begin{aligned} |Y| &= |O + N| \\ &= \sqrt{|O|^2 + |N|^2 + 2|O||N|} \\ &= \sqrt{|O|^2 + |N|^2}, \end{aligned} \tag{5}$$

where the T-F unit symbols were omitted and $\theta$ denotes the phase difference between clean speech and noisy speech within the T-F unit. Since clean speech and additive noise are considered completely independent, the $\theta$ values are random. So the value range of $\cos\theta$ is $[-1, 1]$. Hence, we set $\cos\theta$ to its mean value zero in (5), which should not introduce too much error.

Consequently, the training target of the IRM method is usually expressed as follows:

$$M^{\text{irm}}(m, f) = \frac{|O(m, f)|}{\sqrt{|O(m, f)|^2 + |N(m, f)|^2}}. \tag{6}$$

Sigmoid is used as the output activation function. The amplitude of the target clean speech is obtained by multiplying the amplitude of the corrupted speech with the estimated. For reverberant speech enhancement, this paper replaces additive noise with the difference between reverberant speech and clean speech, which ensures that the target clean speech of the IRM method is the original clean speech, and the expression is

$$M^{\text{irm}}(m, f) = \frac{|O(m, f)|}{\sqrt{|O(m, f)|^2 + |Y(m, f) - O(m, f)|^2}}. \tag{7}$$

Here, $Y$ mainly contains reverberant speech.

The speech enhancement methods studied in this paper only enhance the amplitude of the corrupted speech without dealing with the corrupted phase. Clean speech in the time domain is recovered with corrupted phase and enhanced amplitude as shown in Figure 1.

## 3. Amplitude Fusion Methods

Supervised speech enhancement in this paper refers to training a DNN with training corpus and then converting the corrupted speech into underlying clean speech with the trained DNN. The supervised speech enhancement scheme for the amplitude fusion method is shown in Figure 1. The input features of the DNN can be various extracted speech features [30–32], but the LPS of speech is the most frequently used. The types of DNN here include feedforward multilayer perceptron (MLP) [12], convolutional neural network (CNN) [33, 34], recurrent neural network (RNN) [23, 35], and hybrid neural network [36–38], of which MLP is the most classic neural network. The expected output here refers specifically to the mapping or masking based training target. Based on different conversion mechanisms, the fusions between the mapping and masking methods are proposed in this paper for speech enhancement. We use multitarget training to estimate the training targets; as shown in Figure 1, the selected content in the dashed box manifests the method of multitarget training. "LMS (mapping)" and "LMS (clean)" refer to the estimated LMS by mapping method and the LMS of original clean speech, respectively. $\widehat{M}^{\text{mask}}$ and $M^{\text{mask}}$ are the estimated and the reference masking training target. We use MLP as the neural network, with mapping and masking sharing the weights before the output layer, and the loss function is

$$\begin{aligned} E = \sum_{m,f} \Big[ & \alpha \Big( \widehat{Z}^{\text{map}}(m, f) - Z^{\text{map}}(m, f) \Big)^2 \\ & + (1 - \alpha) \Big( \widehat{M}^{\text{mask}}(m, f) - M^{\text{mask}}(m, f) \Big)^2 \Big], \end{aligned} \tag{8}$$

where $\widehat{Z}^{\text{map}}$ and $Z^{\text{map}}$ are the estimated and the original clean LMS, respectively. $0 < \alpha < 1$, and $\alpha$ refers to the weight coefficient of the two error items. One multitarget training can reduce computational complexity without reducing the performance of multiple single-target pieces of training. For the fusion modes of amplitudes, in addition to the geometric mean (GM) already used in [23], this paper also proposes the arithmetic mean (AM) and the weighted mean (WM) as follows:

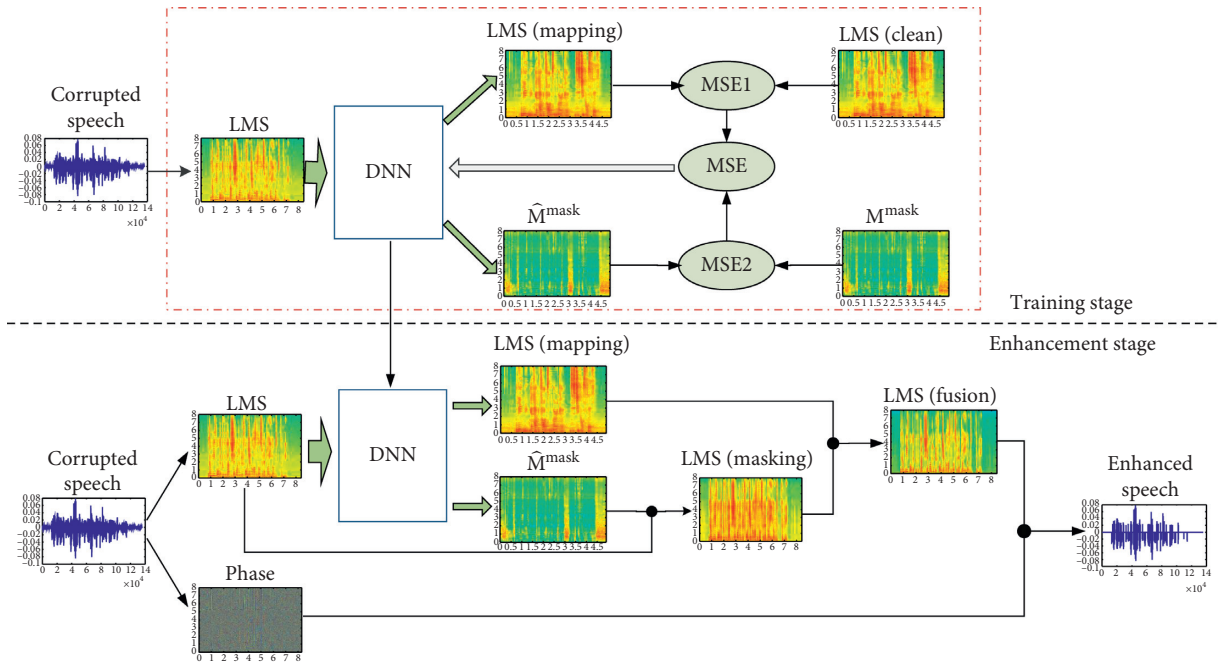


Figure 1: Supervised speech enhancement structure diagram based on the fusion method.

Geometric mean is

$$\widehat{A}^{f} = \sqrt{\widehat{A}^{map} \cdot \widehat{A}^{mask}}. \tag{9}$$

Arithmetic mean is

$$\widehat{A}^{f} = \frac{1}{2}\left(\widehat{A}^{map} + \widehat{A}^{mask}\right). \tag{10}$$

Weighted mean is

$$\widehat{A}^{f} = \beta \cdot \widehat{A}^{map} + (1-\beta) \cdot \widehat{A}^{mask}. \tag{11}$$

LMS-based weighted mean (LWM) is

$$\widehat{Z}^{f} = \gamma \cdot \widehat{Z}^{map} + (1-\gamma) \cdot \widehat{Z}^{mask}. \tag{12}$$

Here, $\widehat{A}^{f}$, $\widehat{A}^{map}$, and $\widehat{A}^{mask}$ refer to the estimated amplitudes of the fusion, mapping, and masking methods, respectively. $\widehat{Z}^{mask}$ is the estimated LMS by masking method. GM, AM, and WM are amplitude-based mean modes, while LWM is an LMS-based weighted mean mode. $\beta$ and $\gamma$ refer to the weight coefficients of the mapping method. They are not some fixed interpolation weights as in [25] but are determined by the following equations:

$$\beta(m,f) = \frac{A(m,f) - \widehat{A}^{mask}(m,f)}{\widehat{A}^{map}(m,f) - \widehat{A}^{mask}(m,f)}, \tag{13}$$

$$\gamma(m,f) = \frac{Z(m,f) - \widehat{Z}^{mask}(m,f)}{\widehat{Z}^{map}(\text{map}) - \widehat{Z}^{mask}(m,f)}, \tag{14}$$

$$\beta = \min(\max(\beta, 0), 1), \tag{15}$$

$$\gamma = \min(\max(\gamma, 0), 1), \tag{16}$$

where $A$ and $Z$ refer to the original clean speech amplitude and LMS. "min" denotes an operation that picks the minimum value of its two parameters by element, and "max" denotes the maximum value. The weight coefficients $\beta$ and $\gamma$ are all obtained by multitarget training as shown in Figure 2. Different from the conventional training methods, $\beta$ and $\gamma$ are not directly labeled by the training set corpus but are calculated by the amplitudes of mapping and masking obtained through the inference process of the neural network. The calculation is performed by (13)–(16). Thereafter, the calculated $\beta$ or $\gamma$ is used as the label value of the weight coefficient in the training process. During the DNN training stage based on the weighted mean mode, the inference process and the training process are performed in each batch. The inference process does not change the parameters of the DNN, but only to calculate the weight coefficient $\beta$ or $\gamma$. The loss function of the training process is as follows:

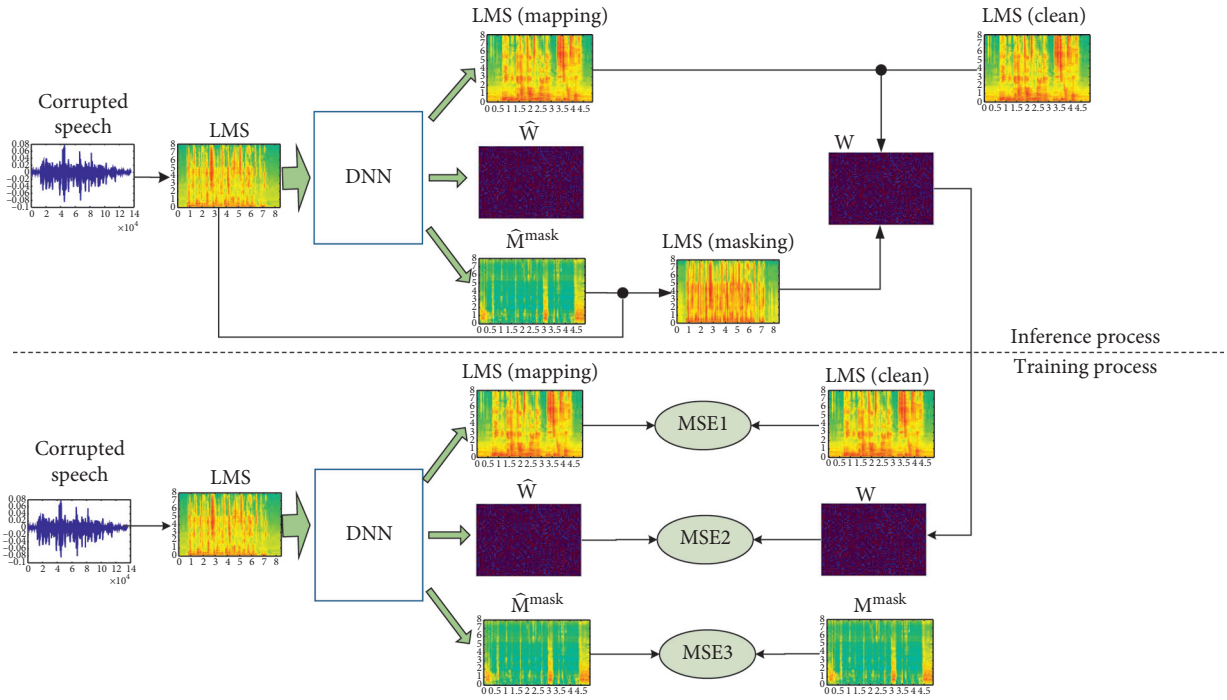


Figure 2: Multitarget training scheme based on weighted mean fusion.

$$E = \sum_{m,f}\Big[\zeta^1 \cdot \left(\widehat{Z}^{map}(m,f) - Z^{map}(m,f)\right)^2 + \zeta^2\left(\widehat{M}^{mask}(m,f) - M^{mask}(m,f)\right)^2 + \zeta^3 \cdot \left(\widehat{W}(m,f) - W(m,f)\right)^2\Big]. \quad (17)$$

Here, $\zeta^1$, $\zeta^2$, and $\zeta^3$ are the weighting coefficients of the error items. $\widehat{W}$ and $W$ represent the estimated and the reference $\beta$ or $\gamma$, respectively. The weighted mean mode ensures that the amplitude obtained by the fusion method is closest to the clean speech amplitude. In fact, when the weight coefficients defined by formulas (11)–(16) are accurately estimated (only in the ideal case), the weighted mean value can achieve zero error compared with the ideal value. Since the amplitude distributions obtained by the mapping and masking methods are regular, the derived $\beta$ or $\gamma$ is also regular. Therefore, the correspondence between them can be learned through the neural network.

## 4. Analysis

*4.1. Conversion Mechanism.* During the training stage of the MAP method, multiple speeches with varying degrees of corruption will be recovered to the same original clean speech. Also during the speech enhancement process, the same clean speech feature is obtained under ideal conditions when enhancing different corrupted speeches originating from the same clean speech. We call this conversion mechanism of MAP "many-to-one." According to the definition of the training target of masking, it is easy to see that the masking method is a "one-to-one" conversion. In order to manifest the relationship between these two conversion mechanisms more clearly, this paper proposes a new training target: the difference of LMS between corrupted speech and clean speech (DCC). DCC is essentially a masking method. In fact, it can also be considered as a variant of IAM. Its expression is as follows:

$$M^{dcc}(m,f) = \log(|Y(m,f)|) - \log(|O(m,f)|) = \log\left(\frac{|Y(m,f)|}{|O(m,f)|}\right). \quad (18)$$

The value of $M^{dcc}$ locates mostly in $(-5,5)$ within our preliminary experiments, and the output activation function is linear. The estimated amplitude of the clean speech is obtained by the following formula:

$$|\widehat{O}(m,f)| = \exp\left(\log(|Y(m,f)|) - \widehat{M}^{dcc}(m,f)\right). \quad (19)$$

Here, $\widehat{O}$ and $\widehat{M}^{dcc}$ refer to the estimated original clean speech and the estimated $M^{dcc}$, respectively. The function $\exp(\cdot)$ denotes the exponential expansion. From (18), the following relational equation can be easily obtained:

$$\log(|Y(m,f)|) = \log(|O(m,f)|) + M^{dcc}(m,f). \quad (20)$$

The left side of (20) is the corrupted speech's LMS which is usually used as the input feature for the mapping or masking methods. The first item on the right side is the training target of the mapping method, and the second item is the training target of the masking method. DCC has quite a few interesting properties. (1) It extends the physical meaning of masking method, so that the training target of masking can be understood not only as a ratio but also as a difference. (2) It helps to establish the relationship between mapping and masking method, so that the mapping and

masking training targets can be regarded as a decomposition of corrupted speech. (3) The DCC training target has a logarithmic function structure similar to the MAP training target, which is beneficial for us to analyze the error minimization mechanism of amplitude fusion method in this paper.

*4.2. Error Minimization.* As mentioned in [11], let $\eta$ mark the ratio between the estimated amplitude and the clean speech amplitude, and $\eta \in [0, +\infty)$. The specific expression of $\eta$ is as follows:

$$\eta = \frac{\widehat{A}}{A}. \tag{21}$$

Here, $\widehat{A}$ refers to the estimated amplitude. When $\eta > 1$ or when $\eta < 1$, the method overestimates or underestimates the amplitude, respectively. Of course, the method get the minimum prediction error when $\eta = 1$. The estimated amplitude of mapping, masking, and the fusion method can be expressed as

$$\widehat{A}^{\text{map}} = A \cdot \eta^{\text{map}}, \tag{22}$$

$$\widehat{A}^{\text{mask}} = A \cdot \eta^{\text{mask}}, \tag{23}$$

$$\widehat{A}^{f} = A \cdot \sqrt{\eta^{\text{map}} \cdot \eta^{\text{mask}}}, \tag{24}$$

where $\eta^{\text{map}}$ and $\eta^{\text{mask}}$ refer to $\eta$ with mapping and masking methods, respectively. Due to the different conversion mechanisms of mapping and masking methods, the resulting $\eta$ values also tend to be different distributions. As an example, when $\eta^{\text{map}} < 1$, it is likely that $\eta^{\text{mask}} > 1$. We call this phenomenon complementarity. The fusion amplitude in (24) is the geometric mean of the two predicted amplitudes [23]; hence the recovery error will be reduced. Here, we only analyze the geometric mean. In fact, the arithmetic mean and weighted mean can also reduce the amplitude recovery error.

*4.3. Proposed Metric on Recovered Amplitudes.* This paper proposes the standard deviation $\sigma$ of $\eta$ as follows:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (\eta_i - \overline{\eta})^2}. \tag{25}$$

In (25), $n$ refers to the total number of speech amplitudes, and $i$ indexes the amplitudes. $\overline{\eta}$ refers to the average of all $\eta$. It is widely shared that the closer the $\eta$ value is to 1, the smaller the prediction error [11], but the fact is not strictly like this. We know that the STFT and the inverse STFT for speech signal have linear property, as shown in the following formulas:

$$F(\lambda \cdot y(t)) = \lambda \cdot Y(f), \tag{26}$$

$$F^{-1}(\lambda \cdot Y(f)) = \lambda \cdot y(t). \tag{27}$$

Here, $F$ and $F^{-1}$ refer to the STFT operation and its inverse operation, respectively. $\lambda$ represents any constant. Since it does not substantially change the speech quality adding a multiplier to a speech signal, the ability of $\eta$ to characterize the prediction accuracy is very limited, while the standard deviation $\sigma$ reflects the degree of dispersion of $\eta$. The smaller the value of $\sigma$, the higher the consistency between the recovered amplitudes and the clean speech amplitudes. In the ideal case $\sigma = 0$, the quality of the recovered amplitudes reaches the upper limit. As can be seen from (25), $\sigma = 0$ when $\eta$ takes any fixed value. Therefore, $\sigma$ is more responsive to the quality of the recovered amplitudes than $\eta$.

## 5. Experiments

*5.1. Experiment Setup.* This paper used REVERB (REverberant Voice Enhancement and Recognition Benchmark) challenge corpus [39] to evaluate the new methods. Its training set for single-channel speech enhancement consists of 7861 synthesized reverberant utterances and the same number of parallel clean utterances. Reverberant utterances are simulated by convolving WSJCAM0 corpus [40] with recorded RIRs. These 7861 clean utterances are different from each other, and there are a total of 24 RIRs recorded with different reverberant conditions. Each RIR is convolved with 328 (or 327) clean utterances to form the reverberant utterances. Evaluation set and development set have similar corpus structure. RIRs are from 6 different reverberant conditions: 3 rooms with different volumes ($S$ = small, $M$ = medium, and $L$ = large); 2 types of distances between a speaker and a microphone (near = 50 cm and far = 200 cm), and the evaluation set or develop set contains 2176 different clean utterances. Each RIR is convolved with 363 (or 362) clean utterances to form the reverberant utterances. The speech or RIRs used in the evaluation set, development set, and training set are all different from each other.

The DNN used in this paper was the feedforward multilayer perceptron (MLP), with three hidden layers, and the number of nodes in each hidden layer was 3072. The activation function of hidden layers was Rectified Linear Unit (ReLU) [41], while the output activation function was linear except sigmoid for IRM. We added a batch normalization layer [42] before the three hidden layers to enhance the generalization ability of MLP to the dynamic features. The mean square error (MSE) was used as the loss function, and $a$ for bitarget training was set to 0.5. For tritarget training based on the weighted mean, $\zeta^1$, $\zeta^2$, and $\zeta^3$ are set to 1. MLP was trained with Adam optimizer [43] and a learning rate of 0.0002, using a batch size of 200. The maximum number of training epochs was set to 80. Once the network was trained, the model with the smallest recovery error on development dataset was chosen and testing on evaluation dataset was performed. The parameter settings for the DNN are listed in Table 1.

For speech segmentation, the frame length was 32 ms, and the frameshift was 16 ms, while the window type was Hanning window. The input feature was the reverberant speech's LMS which was normalized to zero mean and unit

Table 1: Parameters setting up for DNN.

| Parameter | DNN |
|---|---|
| Type | MLP |
| Number of hidden layers | 3 |
| Number of nodes in hidden layer | 3072 |
| Input feature | LMS |
| Input context | Contextual 7 frames |
| Output context | Contextual 7 frames |
| Activation function for hidden layer | ReLU |
| Activation function for output layer | Linear for MAP, IAM, and DCC sigmoid for IRM |
| Loss function | MSE |
| Optimizer | Adam |
| Learning rate | 0.0002 |
| Batch size | 200 |
| Number of training epochs | 80 |

variance. The parameters of the speech analysis are listed in Table 2.

For a single-target prediction method, each input or output feature of the MLP contains 7 frames (left 3 and right 3) context in order to utilize the contextual information. As the sliding distance of the input or output feature is one frame, each frame in the reverberant speech is enhanced 7 times to get 7 predicted amplitudes. Multiple predictions of a frame in an utterance are averaged; that is, the ultimate recovered amplitudes were obtained by averaging the 7 predicted amplitude values of the same frame. The specific expressions for the ultimate recovered amplitudes of the various methods are as follows:

$$\widehat{A}^{\text{map}} = \frac{1}{7}\sum_{i=1}^{7}\widehat{A}_i^{\text{map}}, \tag{28}$$

$$\widehat{A}^{\text{mask}} = \frac{1}{7}\sum_{i=1}^{7}\widehat{A}_i^{\text{mask}}, \tag{29}$$

$$\widehat{A}^{f} = \sqrt{\widehat{A}^{\text{map}} \cdot \widehat{A}^{\text{mask}}}, \tag{30}$$

where $i$ indexes the output batches of the same speech frame during the speech enhancement stage. The evaluation of speech quality is based on the perceptual evaluation of speech quality (PESQ) [44] and the frequency-weighted segmental signal-to-noise ratio (SNR) [45].

*5.2. Experiment Results.* In addition to the various original mapping and masking methods, our comparison methods also include their important variants and fusion methods. The variants and their descriptions are listed in Table 3.

The fusion methods tested in this paper include MAP + I_IRM, MAP + DCC, MAP + IAM, MAP + IRM, IAM + IRM, IAM + DCC, and IRM + DCC, where "A" and "B" in "$A + B$" represent the different single-target prediction methods and "$A + B$" refers to the fusion method of "A" and "B."

*5.2.1. Evaluation of Different Speech Enhancement Methods.* The test results of different speech enhancement methods on the REVERB dataset are shown in Table 4. Without loss of generality, the fusion methods in this table are based on the geometric mean mode. MIX in the table refers to the unprocessed reverberant speech. The scores in Table 4 are averages over the entire evaluation dataset. It is shown that there is a big gap between the WPE method and the supervised speech enhancement methods. Although IRM and I_IRM perform well in speech denoising, they work poorly on speech dereverberation. The PSM and cIRM methods that consider both phase and amplitude did not exceed the IAM method that only enhances amplitude. This may be because the phase wrapping problem is inherent and does not disappear with the change of its expression. TDR exceeds IAM in SNR but is significantly lower than IAM in PESQ. We think this is related to the low recognizability of the time-domain signal. The other variants of the masking methods including I_IRM, IAM_A, and DCC_A did not exceed their original forms in our corpus. MAP + I_IRM was proposed in [23] for removing additive noise, but it is not good enough at speech dereverberation. We choose the traditional masking methods with high scores for amplitude fusion research. The fusion methods MAP + DCC and MAP + IAM have a significant improvement in both speech metrics compared to their single-target prediction methods. Although MAP + IRM is lower in SNR than MAP, the improvement on PESQ is very significant. The MAP + DCC method achieves the best results among all methods. In summary, the amplitude fusion of the mapping and masking methods has significant improvement, and Table 4 shows that the scores of IRM + DCC, IAM + IRM, or IAM + DCC are not improved compared to their single-target prediction methods, only the average of the two. This shows that the amplitude fusion between masking training targets cannot further improve the speech quality.

Table 2: Analysis conditions for the features.

| Parameter | LMS |
|---|---|
| Sampling rate | 16000 |
| Window type | Hanning |
| Frame length | 32 ms |
| Frame shift | 16 ms |
| FFT size | 512 |
| Dimension | 257 |

Table 3: Description of some important methods derived from masking.

| Method | Basic principle |
|---|---|
| TDR | Time-domain signal reconstruction. This paper uses IAM-based TDR, and also clean speech phase is used to recover the time-domain signal [46–48]. |
| I_IRM | Indirect mapping of IRM, which was proposed in [23] to learn the IRM target via MSE between the masked and reference clean LMS. |
| IAM_A | In this method, the DNN estimates a IAM mask that is applied over the corrupted speech amplitude and the loss function is created between masked amplitude and the clean speech amplitude [49, 50]. |
| DCC_A | This method is similar to IAM_A, except that IAM mask is replaced with DCC mask. |

Table 4: Average scores of various speech enhancement methods.

| Type | Method | PESQ | SNR |
|---|---|---|---|
| — | MIX | 1.48 | 3.62 |
| — | WPE | 1.60 | 4.23 |
| Mapping | MAP | 1.77 | 7.59 |
| Masking | cIRM | 1.63 | 7.18 |
| | PSM | 1.80 | 6.93 |
| | TDR | 1.77 | 7.76 |
| | I_IRM | 1.54 | 4.02 |
| | IAM_A | 1.85 | 7.71 |
| | DCC_A | 1.83 | 7.64 |
| | IRM | 1.57 | 4.24 |
| | IAM | 1.91 | 7.17 |
| | DCC | 1.86 | 8.21 |
| Masking + masking | IAM + IRM | 1.74 | 5.91 |
| | IRM + DCC | 1.79 | 6.66 |
| | IAM + DCC | 1.90 | 7.89 |
| Mapping + masking | MAP + I_IRM | 1.90 | 7.30 |
| | MAP + IRM | 1.91 | 7.14 |
| | MAP + IAM | 2.00 | 8.19 |
| | MAP + DCC | **2.01** | **8.78** |

*5.2.2. Performance Analysis under Different Levels of Reverberation.* In order to compare the performance under different levels of reverberation, the PESQ and SNR scores were analyzed for the main methods in different sizes of rooms and different recording distances as shown in Tables 5 and 6. AVE at the bottom represents the average scores of the various methods. Table 5 shows that the MAP method performs poorly in small rooms, especially in the near field, while in medium and large rooms it performs very well. Masking methods, such as IAM, IRM, and DCC, can effectively improve speech quality under a variety of reverberant conditions, although IRM improves very little. It is not difficult to find that the mapping and masking methods are complementary from the perspective of speech quality under various reverberant conditions. The fusion of MAP and masking methods has greatly improved the speech quality. Among them, MAP + DCC greatly improves the speech quality under various degrees of reverberation conditions. Although slightly lower than IAM in small rooms, MAP + IAM is significantly improved under other reverberation conditions. In addition to the small reduction in the medium and large rooms in far field, there is a great improvement with MAP + IRM under other reverberant conditions. However, Table 5 shows that the fusion between different masking methods does not effectively improve the speech quality under any reverberant conditions.

As can be seen from Table 6, MAP + IRM does not improve the SNR score when the reverberation is severe, which may be caused by the traditional IRM method not suitable for speech dereverberation. Although the MAP + IAM score improves on average compared with MAP and IAM, it does not improve in the far field or the small room. MAP + DCC not only has an average score increase but also improves significantly under all conditions of severe reverberation. This shows a high degree of complementarity between MAP and DCC. It may be caused by the same compression function natural logarithm of MAP and DCC. The log makes their predicted amplitudes have the same margin of error, which helps the errors cancel each other out. For the phenomenon that the scores of MAP + IAM and MAP + DCC decrease in the small room, we think that this is caused by the margin of amplitude error of MAP being too large relative to the masking method.

In order to analyze the fusion method more thoroughly, the PESQ and SNR scores are compared again according to different SNRs. The comparison results are listed in Tables 7 and 8, respectively. The leftmost column in the table is the average SNR score of unprocessed reverberant speech under different reverberation conditions and is arranged according to size. Through the comparison of MAP and IAM, it is found that MAP and IAM are complementary in different SNRs, which is reflected in both PESQ and SNR. Specifically,

Table 5: PESQ scores of various methods under different levels of reverberation.

| Type | | — | Mapping | Masking | | | Masking + masking | | | Mapping + masking | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | | MIX | MAP | IRM | IAM | DCC | IAM + IRM | IRM + DCC | IAM + DCC | MAP + IRM | MAP + IAM | MAP + DCC |
| Far field | S | 1.61 | 1.82 | 1.71 | 2.13 | 2.07 | 1.91 | 1.97 | 2.14 | 1.99 | 2.10 | 2.12 |
| | M | 1.19 | 1.49 | 1.25 | 1.47 | 1.51 | 1.35 | 1.39 | 1.51 | 1.45 | 1.57 | 1.59 |
| | L | 1.17 | 1.45 | 1.22 | 1.44 | 1.48 | 1.32 | 1.35 | 1.47 | 1.38 | 1.51 | 1.53 |
| Near field | S | 2.14 | 2.08 | 2.26 | 2.74 | 2.51 | 2.52 | 2.58 | 2.62 | 2.62 | 2.62 | 2.59 |
| | M | 1.40 | 1.91 | 1.52 | 1.85 | 1.78 | 1.70 | 1.75 | 2.06 | 2.06 | 2.14 | 2.13 |
| | L | 1.37 | 1.89 | 1.47 | 1.85 | 1.82 | 1.67 | 1.72 | 1.95 | 1.95 | 2.06 | 2.08 |
| AVE | | 1.48 | 1.77 | 1.57 | 1.91 | 1.86 | 1.74 | 1.79 | 1.91 | 1.91 | 2.00 | 2.01 |

Table 6: SNR scores of various methods under different levels of reverberation.

| Type | | — | Mapping | Masking | | | Masking + masking | | | Mapping + masking | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | | MIX | MAP | IRM | IAM | DCC | IAM + IRM | IRM + DCC | IAM + DCC | MAP + IRM | MAP + IAM | MAP + DCC |
| Far field | S | 6.68 | 9.20 | 7.50 | 11.11 | 11.16 | 9.68 | 10.43 | **11.17** | 9.37 | 10.38 | 10.95 |
| | M | 1.04 | 5.55 | 1.65 | 4.18 | 4.71 | 2.51 | 3.22 | 4.21 | 4.18 | 5.21 | **5.83** |
| | L | 0.24 | 5.49 | 0.82 | 4.05 | 4.85 | 1.88 | 2.64 | 4.12 | 3.46 | 4.77 | **5.79** |
| Near field | S | 8.12 | 10.62 | 8.91 | 13.54 | 13.37 | 11.79 | 12.37 | **13.56** | 11.78 | 12.66 | 12.82 |
| | M | 3.35 | 7.24 | 3.68 | 7.19 | 7.47 | 5.21 | 6.16 | 7.12 | 7.49 | 8.20 | **8.63** |
| | L | 2.27 | 7.47 | 2.86 | 7.15 | 7.71 | 4.36 | 5.16 | 7.14 | 6.57 | 7.90 | **8.67** |
| AVE | | 3.62 | 7.59 | 4.24 | 7.87 | 8.21 | 5.91 | 6.66 | 7.89 | 7.14 | 8.19 | **8.78** |

MAP performs better at lower SNRs, while being lower than IAM at higher SNRs. The comparison of MAP with IRM or DCC also has similar rule. The PESQ and SNR scores of MAP + IRM improve significantly at higher SNR but do not at lower SNR. MAP + IAM improves PESQ when the SNR is low, and its SNR value improves only when the MIX SNR takes the middle value. The PESQ score of MAP + DCC significantly improves at any SNR, and the SNR score improves significantly when the MIX SNR is less than 6.68. It is proved again that MAP and DCC have a higher degree of fusion, and the fusion effect is better at a lower SNR. We conclude that the analytical conclusions obtained under reverberation and noise conditions are consistent.

*5.2.3. Evaluation of Different Fusion Modes.* Without loss of generality, this paper compares various fusion modes based on the MAP and DCC method. The experimental results are shown in Table 9.

The results show that all fusions exceed the traditional MAP or DCC method. Among them, GM and LWM score higher. According to the definition of GM in (9), GM can also be regarded as an arithmetic mean based on LMS. Perhaps the LMS is more capable of characterizing speech signal. We speculate that LWM can get the highest score because it is based on LMS.

*5.2.4. Listening Test.* In addition to the objective evaluation, we also conducted a listening test on the main methods. The corpus used for the test was from the evaluation set of REVERB. Seventeen sentences were randomly selected from the reverberant speech of each condition (far and near fields; small, medium, and large rooms), and a total of 102 sentences were obtained for the listening test. Speech enhancement methods for comparison include MAP, IAM, MAP + I_IRM based on GM mode, and MAP + DCC based on LWM mode. The enhanced speeches are placed under 102 folders, each of which contains four audio files. They all come from the same sentence, only from different algorithms. The order of the four audio files is completely random, and the subjects do not know the order. The five subjects (two males and three females) are all from graduate students or staff at Tianjin University. Their ages are between 24 and 37. English is not their native language. Each participant received a monetary incentive for listening test. They were instructed to compare and sort the four audio in each folder. The content of the comparison is the degree of distortion and completeness of the speech, as well as the effect of suppressing noise and reverberation. The basis for sorting is the listener's own overall feeling of speech quality. The favorite audio ranks first, the second favorite ranks second, the third favorite ranks third, and the dislike ranks fourth. Listeners use a high-quality headset to test in a quiet environment and play each audio at least once. The ranking is scored and the scoring method borrows from [19]. The first is given a score of 3, the second 2, the third 1, and the fourth 0. After the listening test, the total score corresponding to each method is calculated. The score results are shown in Figure 3. In the figure, "m1-2" mark the scores of two males, and "f1-3" three females. AVE marks the average of the scores of the five subjects. The values of AVE corresponding to MAP, IAM, MAP + DCC, and MAP + I_IRM are 133.8, 162.2, 183.0, and 133.0. If the percentages are used for comparison, they account for 21.9%, 26.5%, 29.9%, and 21.7%, respectively. Our proposed MAP + DCC scores the highest score in the listening test.

Table 7: PESQ scores of various methods with respect to the SNRs.

| SNR | PESQ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| — | Mapping | Masking | | | Masking + masking | | | Mapping + masking | | |
| MIX | MAP | IRM | IAM | DCC | IAM + IRM | IRM + DCC | IAM + DCC | MAP + IRM | MAP + IAM | MAP + DCC |
| 8.12 | 2.08 | 2.26 | 2.74 | 2.51 | 2.52 | 2.58 | 2.62 | 2.62 | 2.62 | 2.59 |
| 6.68 | 1.82 | 1.71 | 2.13 | 2.07 | 1.91 | 1.97 | 2.14 | 1.99 | 2.10 | 2.12 |
| 3.35 | 1.91 | 1.52 | 1.85 | 1.78 | 1.70 | 1.75 | 1.82 | 2.06 | 2.14 | 2.13 |
| 2.27 | 1.89 | 1.47 | 1.85 | 1.82 | 1.67 | 1.72 | 1.83 | 1.95 | 2.06 | 2.08 |
| 1.04 | 1.49 | 1.25 | 1.47 | 1.51 | 1.35 | 1.39 | 1.51 | 1.45 | 1.57 | 1.59 |
| 0.24 | 1.45 | 1.22 | 1.44 | 1.48 | 1.32 | 1.35 | 1.47 | 1.38 | 1.51 | 1.53 |
| AVE | 1.77 | 1.57 | 1.91 | 1.86 | 1.74 | 1.79 | 1.90 | 1.91 | 2.00 | 2.01 |

Table 8: SNR scores of various methods with respect to the SNRs.

| — | Mapping | Masking | | | Masking + masking | | | Mapping + masking | | |
|---|---|---|---|---|---|---|---|---|---|---|
| MIX | MAP | IRM | IAM | DCC | IAM + IRM | IRM + DCC | IAM + DCC | MAP + IRM | MAP + IAM | MAP + DCC |
| 8.12 | 10.62 | 8.91 | 13.54 | 13.37 | 11.79 | 12.37 | 13.56 | 11.78 | 12.66 | 12.82 |
| 6.68 | 9.20 | 7.50 | 11.11 | 11.16 | 9.68 | 10.43 | 11.17 | 9.37 | 10.38 | 10.95 |
| 3.35 | 7.24 | 3.68 | 7.19 | 7.47 | 5.21 | 6.16 | 7.12 | 7.49 | 8.20 | 8.63 |
| 2.27 | 7.47 | 2.86 | 7.15 | 7.71 | 4.36 | 5.15 | 7.14 | 6.57 | 7.90 | 8.67 |
| 1.04 | 5.55 | 1.65 | 4.18 | 4.71 | 2.51 | 3.22 | 4.21 | 4.18 | 5.21 | 5.83 |
| 0.24 | 5.49 | 0.82 | 4.05 | 4.85 | 1.88 | 2.64 | 4.12 | 3.46 | 4.77 | 5.79 |
| AVE | 7.59 | 4.24 | 7.87 | 8.21 | 5.91 | 6.66 | 7.89 | 7.14 | 8.19 | 8.78 |

Table 9: Comparison between various fusion modes.

| | MAP | DCC | AM | WM | GM | LWM |
|---|---|---|---|---|---|---|
| PESQ | 1.77 | 1.86 | 1.92 | 1.94 | 2.01 | **2.02** |
| SNR | 7.59 | 8.21 | 8.22 | 8.52 | 8.78 | **8.96** |

*5.3. Discussion.* In terms of speech denoising, we think that MAP + I_IRM may be the most suitable fusion method [23]. However, MAP + DCC is more conducive to speech dereverberation than MAP + I_IRM, which is caused by the excellent performance of DCC on dereverberation. Since the DCC and MAP methods use the same compression function, the complementarity between them is most significant, and the dereverberation effect is also the best.

*5.3.1. Issue on the Conversion Mechanisms.* Based on the evaluation set corpus, we calculated the $\sigma$ values of MAP, DCC, and their geometric mean fusion MAP + DCC. The results are shown in Table 10.

The masking method as a "one-to-one" conversion is reflected in the amplitude recovery accuracy shown in Table 10. The reverberation time corresponding to the small, medium, and large rooms used in the evaluation set corpus is 0.25s, 0.5s, and 0.7s, respectively. $\sigma$ of the DCC method shows obvious regularity; that is, the lighter the speech reverberation is, the smaller the $\sigma$ becomes. As can be seen from (18), $M^{\text{dcc}}$ contains the amplitude of the reverberant speech. Therefore, the degree of chaos in DCC training target is positively correlated with the degree of speech reverberation. For the same type of prediction target based on deep learning, there may be a rule that the lower the degree of chaos of the prediction target is, the higher the prediction accuracy becomes. This basically agrees with the PESQ and SNR scores of the DCC method shown in Tables 5 and 6. The distribution of $\sigma$ with the MAP item in Table 10 shows that the speech with a lighter reverberation does not achieve higher prediction accuracy. The MAP method yields similar $\sigma$ values under different reverberations. MAP as a "many-to-one" conversion refers to predicting the same corresponding clean speech feature from multiple reverberant speech features of different degrees. The prediction targets of MAP are all clean speech features, which may result in close prediction accuracy under different reverberation

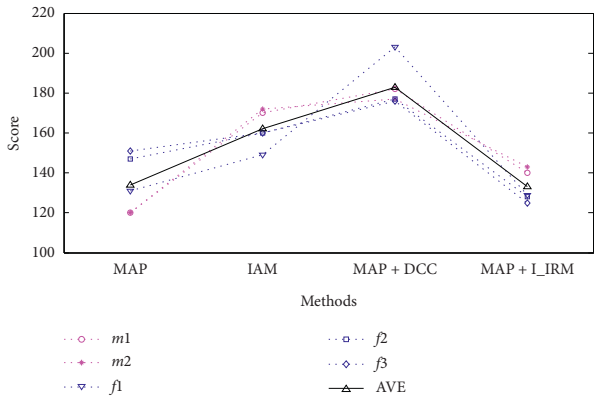


Figure 3: Listening test results.

Table 10: $\sigma$ of various methods under different levels of reverberation.

| | | $\sigma$ | | |
|---|---|---|---|---|
| | | MAP | DCC | MAP + DCC |
| Far field | S | 4.06 | 3.00 | **2.57** |
| | M | 3.75 | 3.74 | **2.86** |
| | L | 3.86 | 4.05 | **2.89** |
| Near field | S | 4.15 | 2.94 | **2.61** |
| | M | 3.87 | 3.18 | **2.66** |
| | L | 3.66 | 4.40 | **3.06** |
| AVE | | 3.89 | 3.55 | **2.78** |

Table 11: $\sigma$ of various methods with respect to the SNRs.

| SNR | $\sigma$ | | |
|---|---|---|---|
| MIX | MAP | DCC | MAP + DCC |
| 8.12 | 4.15 | 2.94 | **2.61** |
| 6.68 | 4.06 | 3.00 | **2.57** |
| 3.35 | 3.87 | 3.18 | **2.66** |
| 2.27 | 3.66 | 4.40 | **3.06** |
| 1.04 | 3.75 | 3.74 | **2.86** |
| 0.24 | 3.86 | 4.05 | **2.89** |
| AVE | 3.89 | 3.55 | **2.78** |

conditions. The phase of the reverberant speech is used to recover target clean speech as shown in Figure 1. The degree of phase chaos is positively correlated with the degree of reverberation of the reverberant speech. This will affect the quality of the recovered speech. Therefore, the amplitude recovery accuracy of the MAP method shown in Table 10 is in good consistent with the PESQ and SNR scores of MAP in Tables 5 and 6.

At the same time, we give the distribution of $\sigma$ values according to different SNRs, which are listed in Table 11. The table shows that the $\sigma$ values of the MAP change slightly under different SNRs, while $\sigma$ of DCC increases as the SNR decreases, except when the SNR is 2.27. MAP + DCC has achieved the minimum value under various SNRs. Therefore, the analysis results of $\sigma$ according to the reverberation and noise conditions are basically consistent.

In summary, we use the different conversion mechanisms of mapping and masking to explain their complementarity in different degrees of speech damage, and this complementarity is also the motivation for the fusion of MAP and I_IRM [23]. From this, we conclude that MAP and other methods belonging to masking are also complementary, so we propose new fusion methods such as MAP + IAM and MAP + DCC.

*5.3.2. Issue on the Error Minimization Mechanism.* In our preliminary experiments, we analyzed $\eta$ produced by different methods on the training set corpus. We divided the training set into 24 subsets according to the RIR. $\eta$ produced by different speech enhancement methods on each subset was compared. We observed significant amplitude error complementarity in 7 of the subsets, as shown in Figure 4.

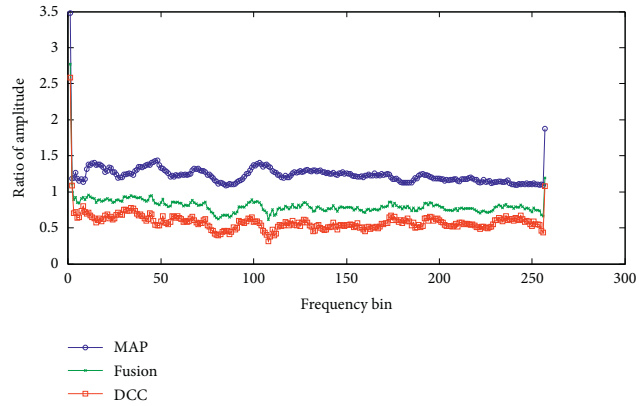


Figure 4: $\eta$ curves of MAP, DCC, and the fusion method.

In the figure, the abscissa value represents the frequency bin, and the ordinate value represents the average value of $\eta$ on one subset. The figure shows that the mean of $\eta$ corresponding to the MAP method is generally greater than 1, while the mean of $\eta$ of DCC is generally less than 1. Therefore, their geometric mean is closer to 1. The 7 corresponding RIRs are recorded on different reverberant conditions.

As shown in Tables 10 and 11, $\sigma$ of the fusion method MAP + DCC is the smallest compared to the other two. This suggests that there is complementarity between the amplitude recovery errors of the MAP and DCC methods, and their geometric mean reduces the errors. Since both DCC and MAP training targets use logarithmic function compression, their amplitude recovery errors have the same scale. Moreover, the existence of complementarity makes it easy to minimize $\sigma$ of the fusion method.

As shown in Tables 4–6, the fusion between the masking methods, such as IRM + DCC, IAM + IRM, and IAM + DCC, does not improve the speech quality. This should be because the recovery errors produced by the same conversion mechanism are less likely to complement each other. However, the fusion between mapping and masking methods, such as MAP + IRM, MAP + IAM, and MAP + DCC, improves significantly. In view of the distribution of $\sigma$ and the scores of speech, it can be inferred that the different conversion mechanisms between the mapping and masking methods are the source of their complementarity.

In theory, the weighted mean method should minimize the prediction error of the amplitude. The experimental results also show that the LWM method achieved the highest score. It can be inferred that LWM will work better if the prediction accuracy of $\gamma$ is improved with a more powerful neural network.

In order to directly compare the speech amplitudes recovered by various methods, we provide their spectrograms. Figure 5 shows that the fusion method is better than other three methods in reducing reverberation. This is in agreement with Table 4, which shows that the MAP + DCC in GM fusion mode achieves the highest speech quality.

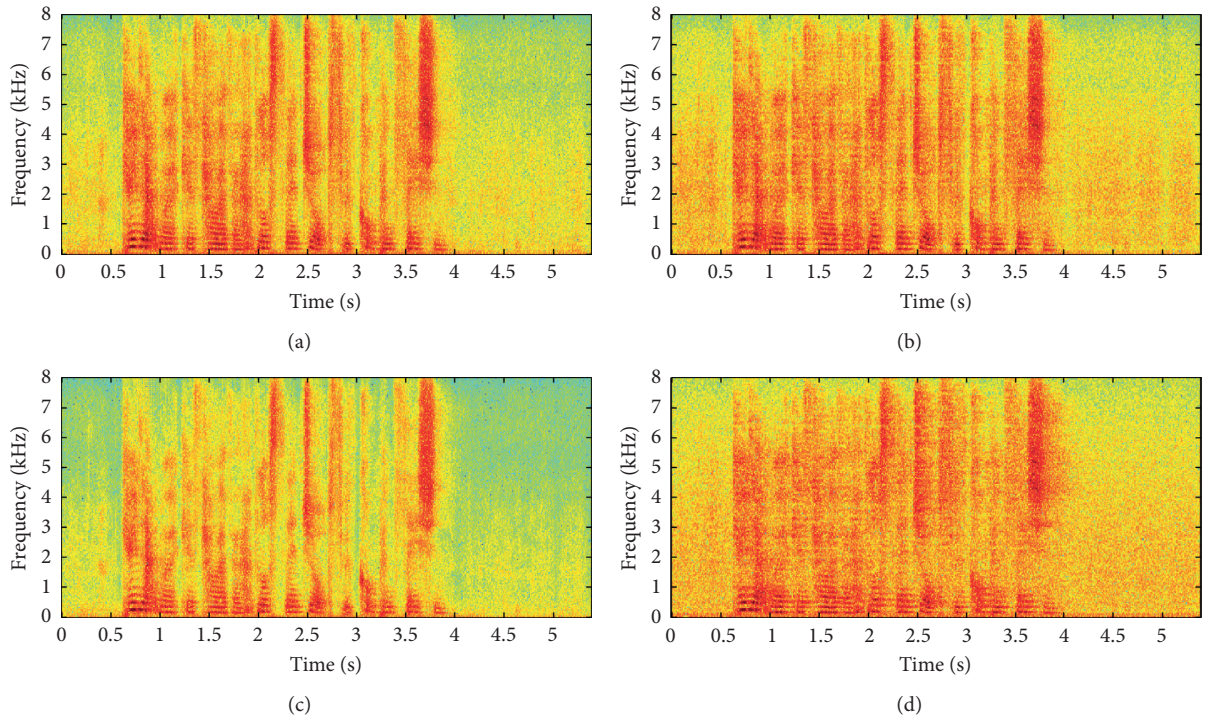


Figure 5: The spectrums of speeches enhanced by MAP, DCC, and MAP + DCC methods and the spectrum of reverberant speech. (a) MAP. (b) DCC. (c) MAP + DCC. (d) Reverberant.

## 6. Conclusion

This paper analyzes in detail the different conversion mechanisms of mapping and masking methods and experimentally verifies that their recovery errors are highly complementary. The amplitude fusion method used in this paper can effectively utilize this complementarity and reduce the recovery error of the target clean speech amplitude, thus further improving the speech quality. Furthermore, it is found that not only can MAP + IRM further improve the speech, but also the fusion of mapping with other masking methods, such as MAP + IAM and MAP + DCC, can greatly improve the speech quality. This is because there is complementarity between the recovery errors due to different conversion mechanisms between mapping and masking methods. Since the masking methods have the same conversion mechanism, the amplitude fusion between them does not further improve the speech quality. This paper proposes a new fusion mode, LWM, and also proposes a new method of predicting weight coefficient with DNN. The predicted target $\beta$ or $\gamma$ is different from the traditional target directly marked through corpus (such as mapping or masking) but obtained by the other two outputs of DNN. LWM takes full advantage of the DNN's predictive power to minimize amplitude prediction errors. This paper also proposes a new method DCC that can be seen as a variant of IAM, and the MAP + DCC fusion method achieves the best results for speech dereverberation. Experiments indicate that the MAP + DCC improved PESQ and SNR by 5.8% and 25.0%, respectively, compared with the traditional IAM method.

## Data Availability

The data used to support this study can be found at https://reverb2014.dereverberation.org/data.html.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61771333 and the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

## References

[1] K. A. Al-Karawi, A. H. Al-Noori, F. Li, and T. Ritchings, "Automatic speaker recognition system in adverse conditions-implication of noise and reverberation on system performance," *International Journal of Information and Electronics Engineering*, vol. 5, no. 6, pp. 423–427, 2015.

[2] D. Gelbart and N. Morgan, "Double the trouble: handling noise and reverberation in far-field automatic speech recognition," in *Proceedings of the ICSLP*, pp. 2185–2188, Denver, Colorado, September 2002.

[3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[4] E. L. J. George, S. T. Goverts, J. M. Festen, and T. Houtgast, "Measuring the effects of reverberation and noise on sentence

intelligibility for hearing-impaired listeners," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 6, pp. 1429–1439, 2010.

[5] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[6] M. Delcroix, T. Yoshioka, A. Ogawa et al., "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *Proceedings of the 2014 REVERB Workshop*, Florence, Italy, May 2014.

[7] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4628–4632, Florence, Italy, May 2014.

[8] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2013.

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

[10] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

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

[12] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[13] K. Han and D. Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[14] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio Speech Lang. Process.*vol. 17, no. 4, pp. 625–638, 2009.

[15] Y. Zhao, D. Wang, I. Merks, and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6525–6529, Shanghai, China, March 2016.

[16] X. Li, J. Li, and Y. Yan, "Ideal ratio mask estimation using deep neural networks for monaural speech segregation in noisy reverberant conditions," in *Proceedings of the INTERSPEECH*, pp. 1203–1207, Stockholm, Sweden, August 2017.

[17] M. Kolbaek, D. Yu, Z.-H. Tan et al., "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[18] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 708–712, Brisbane, Australia, April 2015.

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

[20] D. S. Williamson and D. L. Wang, "Speech dereverberation and denoising using complex ratio masks," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5590–5594, New Orleans, LA, USA, March 2017.

[21] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[22] D. S. Williamson, X. Wang, and D. L. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *in Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224, Shanghai, China, March 2016.

[23] L. Sun, J. Du, L. R. Dai, and C. H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proceedings of the Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pp. 136–140, San Francisco, CA, USA, March 2017.

[24] H. Zhang, X. Zhang, and G. Gao, "Multi-target ensemble learning for monaural speech separation," in *Proceedings of the INTERSPEECH*, pp. 1958–1962, Stockholm, Sweden, August 2017.

[25] B. Li and K. C. Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 279–284, Olomouc, Czech Republic, December 2013.

[26] P. Pertilä, "Data-dependent ensemble of magnitude spectrum predictions for single channel speech enhancement," in *Proceedings of the IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, Kuala Lumpur, Malaysia, September 2019.

[27] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 967–977, 2016.

[28] Z. Q. Wang and D. L. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, p. 1, 2020.

[29] Y. Wang and D. L. Wang, "A structure-preserving training target for supervised speech separation," in *Proceedings of the ICASSP*, pp. 6107–6111, Florence, Italy, May 2014.

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

[31] J. Chen, Y. Wang, and D. Wang, "A feature study for classification based speech separation at very low signal-to-noise ratio," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7059–7063, Florence, Italy, May 2014.

[32] M. Delfarah and D. L. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, 2017.

[33] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," 2018, https://arxiv.org/abs/1803.08243.

[34] D. S. Wang, Y. X. Zou, and W. Shi, "A deep convolutional encoder decoder model for robust speech dereverberation," in *Proceedings of the International Conference on Digital Signal Processing (DSP)*, pp. 1–5, London, UK, August 2017.

[35] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[36] H. Zhao, S. Zarar, I. Tashev, and C. H. Lee, "Convolutional recurrent neural networks for speech enhancement," in *Proceedings of the ICASSP*, pp. 2401–2405, Calgary, Canada, April 2018.

[37] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proceedings of the ICASSP*, pp. 5039–5043, Calgary, Canada, April 2018.

[38] C. X. Li, L. Zhu, S. Xu, P. Gao, and Bo Xu, "CBLDNN-based speakerindependent speech separation via generative adeversial training," in *Proceedings of the ICASSP*, pp. 711–715, Calgary, Canada, April 2018.

[39] K. Kinoshita, M. Delcroix, S. Gannot et al., "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, 2016.

[40] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition," in *Proceedings of the ICASSP*, pp. 81–84, Detroit, MI, USA, May 1995.

[41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of theInternational Conference on Machine Learning*, pp. 808–814, Haifa, Israel, June 2010.

[42] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of theInternational Conference on Machine Learning*, pp. 448–456, Lille, France, July 2015.

[43] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," https://arxiv.org/abs/1412.6980.

[44] International Telecommunication Union, *P. 862.2: Wideband Extension to Recommendation P. 862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, International Telecommunication Union, Geneva, Switzerland, 2007.

[45] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[46] X. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proceedings of the ICASSP*, pp. 4390–4394, Brisbane, Australia, April 2015.

[47] Y. Zhao, Z. Q. Wang, and D. L. Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *Proceedings of the ICASSP*, pp. 5580–5584, New Orleans, LA, USA, March 2017.

[48] Y. Zhao, Z. Q. Wang, and D. L. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 53–62, 2018.

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

[50] F. Weninger, H. Erdogan, S. Watanabe et al., "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, Springer, Liberec, Czech Republic, pp. 91–99, August 2015.