Chinese Journal of Electronics
Vol.32, No.4, July 2023

Teacher-Student Training Approach Using an Adaptive Gain Mask for LSTM-Based Speech Enhancement in the Airborne Noise Environment

HUANG Ping and WU Yafeng

(*School of Power and Energy, Northwestern Polytechnical University, Xi'an 710072, China*)

Abstract — Research on speech enhancement algorithms in the airborne environment is of great significance to the security of airborne systems. Recently, the research focus of speech enhancement has turned from conventional unsupervised algorithms, like the log minimum mean square error estimator (log-MMSE), to the state-of-the-art masking-based long short-term memory (LSTM) method. However, each method has its characteristics and limitations, so they cannot always handle noise well. Besides, the requirements of clean speech and noise data for training a supervised speech enhancement model are difficult to satisfy in the real-world airborne environment. Therefore, in this work, to fully utilize the advantages of those two different methods without any data restrictions, we propose a novel adaptive gain mask (AGM) based teacher-student training approach for speech enhancement. In our method, the AGM, as a robust learning target for the student model, is devised by incorporating the estimated ideal ratio mask from the teacher model into the procedure of the log-MMSE approach. To get an appropriate tradeoff between the two methods, we adaptively update the AGM using a recursive weighting coefficient. Experiments on the real airborne data show that the proposed AGM-based method outperforms other baselines in terms of all essential objective metrics evaluated in this paper.

Key words — Adaptive ideal mask, Teacher-student learning, Long short-term memory (LSTM), Speech enhancement.

I. Introduction

Airborne noise, a series of non-stationary broadband random pulses, significantly impairs the communication between the cabin and the ground personnel [1]. The single-channel speech enhancement [2]–[6] aims to suppress the complicated background noises and recover clean speech from the observed noisy speech. Most of the early algorithms for speech enhancement are based on signal processing [7]–[9]. Generally, they work in the time-frequency (T-F) domain and estimate a spectral gain, based on the probability of speech signal absence or presence, to remove the additive noise in the noisy speech magnitude spectrum. This real-valued spectral gain is usually referred to as a T-F representation of the speech presence probability because its values smoothly vary from 0 to 1. In other words, the accuracy of the spectral gain is the key point for improving speech enhancement performance. The oldest spectral gain function is the suppression rule of the Wiener-type filter [10], which tries to minimize the mean-square error criterion for speech enhancement. Other commonly used spectral gain-based approaches are the minimum mean square error-based (MMSE) estimator [11] and log minimum mean square error (log-MMSE) estimator [12], all of which have been proposed to implement the short-time spectral amplitude estimation of the speech signal. These MMSE methods are based on statistical models, such as the Gaussian distribution model [13] and non-Gaussian distribution models [14]. Due to adding these statistical theories, the log-MMSE estimator, as a superior version of the MMSE estimator, is thought to perform better than other classical methods in terms of noise reduction and speech preservation. Yet all mentioned algorithms are considered to be unsupervised techniques that cannot deal with non-stationary noises due to relying too much on statistical as-

Manuscript Received Sept. 13, 2022; Accepted Nov. 30, 2022. This work was supported by the Fundamental Research Funds for the Central Universities (D5000210974).

DOI:10.23919/cje.2022.00.307

sumptions.

Recently, inspired by the computational auditory scene analysis (CASA), the time-frequency masking-based deep learning methods have achieved great success in speech enhancement areas [15]. In such enhancement algorithms [16]–[18], all of these T-F masks, such as the ideal binary mask (IBM), the ideal ratio mask (IRM), or the phase sensitive mask (PSM), are adopted as the good learning target for the current learning technologies. Besides, another function of those target masks is as an indicator of the speech signal presence or absence, similar to the role of spectral gain in the aforementioned conventional methods, thus they also could regard as the a posteriori speech presence probability representation in the speech enhancement work [19]. In addition, some past related studies have demonstrated that the IRM-based deep learning method with strong data-driven capability has already achieved a state-of-the-art noise reduction performance [20]–[22]. Specifically, a deep neural network (DNN) was exploited to estimate the IRM and this approach could significantly improve the performance of speech enhancement, owing to the powerful modeling ability of DNN [23], [24]. However, the fully connected DNN cannot model the long-term relationship between the neighboring frames, even if it has sophisticated deep architectures. In [25], the authors proposed a recurrent neural network (RNN) to capture contextual information by using its unique recursive structures between the historical frames and the future frames. Then the long short-term memory (LSTM) recurrent neural network, as an improved model based on RNN, was proposed and successfully used in speech enhancement tasks for obtaining superior listening speech quality [26], [27].

From those above deep learning-based methods, it can be found that using a proper training target can increase the naturalness of the enhanced utterance and remove noise segments as much as possible. However, since many masking-based targets, like IRMs, are generated by the ratio of the clean speech energy to the mixture energy at each T-F unit, those methods are constrained by some prerequisites, such as the clean training data or pure noise. Those constraints for the training data may lead to higher mismatch problems between the training and test conditions, and also increase the difficulty and cost of the speech data sampling step. To relieve those limitations, in [28], the authors proposed a teacher-student learning framework to improve the accuracy of the automatic speech recognition (ASR) task. Specifically, this method defined an improved learning target by combining the advantages of the IRM and the gains in the improved minima controlled recursive averaging algorithm (IMCRA) [29]. By using the teacher-student learning framework, the student model training can be easily scaled to arbitrary data. However, this method is highly computational complexity because the IMCRA approach needs two smoothing operations and minimum tracking iterations for updating the continuous noise spectral. In [30], the researchers use a teacher-student framework to obtain a high precision for voice activity detection (VAD), in which vast unconstrained data can be employed to train the student model.

Based on the above introduction, the conventional log-MMSE approach and masking-based deep learning have presented different strengths and weaknesses in the single-channel speech enhancement task. For the unsupervised method, the log-MMSE method is a statistical estimator to minimize the mean-square error between the log-amplitude spectra of the original and the predicted speech, while it often fails to handle non-stationary noises well due to excessively depending on the mathematical assumptions between speech and noise. For masking-based deep learning technique, it seems to perform well in removing non-stationary background noises. However, it requires clean speech or synthetically noised data for training, and its performance is usually degraded by the mismatch between the training and test scenarios.

On the other hand, it is generally known that relying too much on given pairs of clean speech (target) and noisy (input) inevitably prevents the speech enhancement model from generalizing to real-world applications because a limited number of training data simulated by the publicly available datasets cannot match all unseen sounds. And the collection of the corresponding clean utterances or noise-only data is comparatively costlier than noisy data. Previous studies [28], [30] have demonstrated that the teacher-student learning framework, which has been successfully applied to speech recognition and VAD, can alleviate this data constraint for deep learning. Therefore, our work introduces the teacher-student framework and aims to fully utilize the advantages of the classical log-MMSE spectral amplitude estimator and the deep learning technology to find a proper learning target for improving the long short-term memory based airborne speech enhancement performance with less stereo-data constraint.

For that, in this paper, we propose a new adaptive gain mask (AGM) based teacher-student training method for the LSTM speech enhancement framework in the airborne environment. In our method, the AGM is used as a better learning target for the student model. Specifically, the proposed AGM target is designed by incorporating the estimated IRM from the teacher model into the procedure of the log-MMSE approach. Then, during the calculation of the AGM, we add a self-adaptive weighting function to dynamically bridge the gap between those two approaches for generating a robust AGM estimate. Moreover, at the adaptive weighting stage, an adjustable scale factor is proposed to independently control the value range of the weighting function which can further help the AGM target to have good enhancement performance. Next, to relieve the data limitation problem, we employ a student model to estimate the AGM target which is directly used to reconstruct the waveform. This allows our student model can relax the requirements on clean speech and noise training data. Finally, we explore the speech enhancement performance of the proposed method using LSTM structures.

On the whole, the technical significance of this work is to provide a novel method that can not only improve the performance of speech enhancement in the adverse airborne environment but also reduce dependence on clean and noise training data pairs. The main contributions of this work are summarized as follows:

- The proposed adaptive gain mask target is robust to noise and capable of providing better speech enhancement performance for the whole teacher-student framework.
- The newly derived adaptive weighting operation can help the whole framework to fully merge the advantages of the state-of-the-art deep learning technology and unsupervised log-MMSE algorithm, which can further improve the quality of enhanced speech.
- The student model can obtain substantial improvements by using noisy data instead of exclusively relying on clean speech or noise data.

Experiments on real-world airborne data show that the proposed method yields huge improvements in speech enhancement over the reference methods. Furthermore, the realistic noisy data without the underlying clean speech and background noise can be directly employed to further improve the generalization capacity of our student model to unexpected adverse environments, which means our method generalizes well to real-world applications. In some sense, our work is also of great theoretical guiding significance to improve the enhancement performance in practical applications.

The remainder of this paper is organized as follows. In Section II, we briefly present the IRM-based speech enhancement network and the classical log-MMSE filter as preliminaries. Section III gives an overview of the proposed AGM-based teacher-student method and describes how the adaptive weighting procedure works. Section IV and Section V present the experimental setup, the performance evaluation, and the results, respectively. Finally, we summarize this paper in Section VI.

II. Preliminaries

1. The IRM-based deep learning for speech enhancement

Let the noisy speech signal $y(t)$ be described in the time-frequency (T-F) domain as

$$Y(k,l)=X(k,l)+N(k,l) \tag{1}$$

where $X(k,l)$, $N(k,l)$ and $Y(k,l)$ denote the complex short-time Fourier transform (STFT) coefficients of the desired clean speech, noise, and the corrupted speech signal, respectively. Here, k is the frequency bin index, and l corresponds to the frame index.

The IRM is widely used as a mask to represent the speech-dominant or noise-dominant meta information on each T-F unit. According to the additive noise model in (1), the ideal ratio mask $M_{\rm IRM}(k,l)$ can be expressed as follows:

$$M_{\rm IRM}(k,l)=\frac{X^2(k,l)}{X^2(k,l)+N^2(k,l)} \tag{2}$$

In the existing IRM-based deep learning algorithms for speech enhancement, an estimate of the clean spectrum $\hat{X}(k,l)$ can be obtained from a neural network directly by estimating a T-F mask $\hat{M}_{\rm IRM}(k,l)$:

$$\hat{X}(k,l)=\hat{M}_{\rm IRM}(k,l)\cdot Y(k,l) \tag{3}$$

From [28] and [31], a typical way of training such a masking-based network is to employ the mean squared error (MSE) loss of the form:

$$L_{\rm MSE}=\sum_{k,l}\left\|\hat{M}_{\rm IRM}(k,l)-M_{\rm IRM}(k,l)\right\|_2^2 \tag{4}$$

Then this MSE loss is optimized using the mini-batch back-propagation method. Finally, the estimated $\hat{X}(k,l)$ from this network can be transformed into the enhanced speech in the time domain via the inverse short-time Fourier transform.

2. Log-MMSE approach to speech enhancement

The log-MMSE noise suppression algorithm presented in [12] is briefly summarized in this section. Intuitively, this suppressor optimally estimates the short-time magnitude spectrum of the clean speech via using a spectral gain $G_{\text{log-MMSE}}(k,l)$. During the process of calculating the spectral gain, a key role is played by the a priori signal-to-noise-ratio (SNR) $\xi(k,l)$ and the a posteriori SNR $\gamma(k,l)$, which can be defined as

$$\xi(k,l) = \frac{|X(k,l)|^2}{\lambda_n(k,l)} \tag{5}$$

and

$$\gamma(k,l) = \frac{|Y(k,l)|^2}{\lambda_n(k,l)} \tag{6}$$

respectively, where λ_n is the noise power spectrum.

For estimating λ_n, an unbiased MMSE-based noise estimation algorithm [32] was introduced, in which the estimate $\hat{\lambda}_n$ can be updated based on the noisy utterances:

$$\hat{\lambda}_n(k,l) = \alpha_d \hat{\lambda}_n(k,l-1) + (1-\alpha_d)[P_{\text{UMMSE}} \hat{\lambda}_n(k,l-1) + (1-P_{\text{UMMSE}})|Y|^2] \tag{7}$$

$$P_{\text{UMMSE}} = \{1 + \exp(-a_{sig}(\gamma - c_{sig}))\}^{-1} \tag{8}$$

$$a_{sig} = \frac{\xi_{H_1}}{1+\xi_{H_1}}, \quad c_{sig} = \log\left(\frac{P(H_0)}{P(H_1)}(1+\xi_{H_1})\right)\frac{1+\xi_{H_1}}{\xi_{H_1}} \tag{9}$$

where α_d is a smoothing coefficient, and it is set to 0.8 according to [32]. P_{UMMSE} is the speech presence probability estimated by the unbiased MMSE-based estimator. ξ_{H_1} denotes the fixed optimal a priori SNR value when the speech signal is active. And $P(H_0)$ and $P(H_1)$ are respectively the a priori probability for speech spectra absence and presence. From [32], the worst situation was considered, where both the likelihoods of speech absence and presence became identical, i.e., $P(H_1) = P(H_0) = 0.5$. Then the log-MMSE suppressor gain function can be given by

$$G_{\text{log-MMSE}}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left\{\frac{1}{2}\int_{v(k,l)}^{\infty} \frac{e^{-t}}{t} dt\right\} \tag{10}$$

$$v(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \gamma(k,l) \tag{11}$$

where integral in (10) is known as the exponential integral. Finally, the enhanced speech can be achieved by applying the spectral gain to each spectral component of the observed noisy speech as follows:

$$\hat{X}(k,l) = G_{\text{log-MMSE}}(k,l) \cdot Y(k,l) \tag{12}$$

III. Proposed Method

The overall architecture of the proposed AGM-based teacher-student LSTM training framework is illustrated in Fig.1. It mainly consists of three modules, respectively shown in three different dashed bins in this figure. The first module is the teacher training module, the second module is used to compute the adaptive gain mask, and the third one is the student training module. The dotted lines represent the causal process of getting the ideal AGM, while the solid lines denote the learning process of a single module.

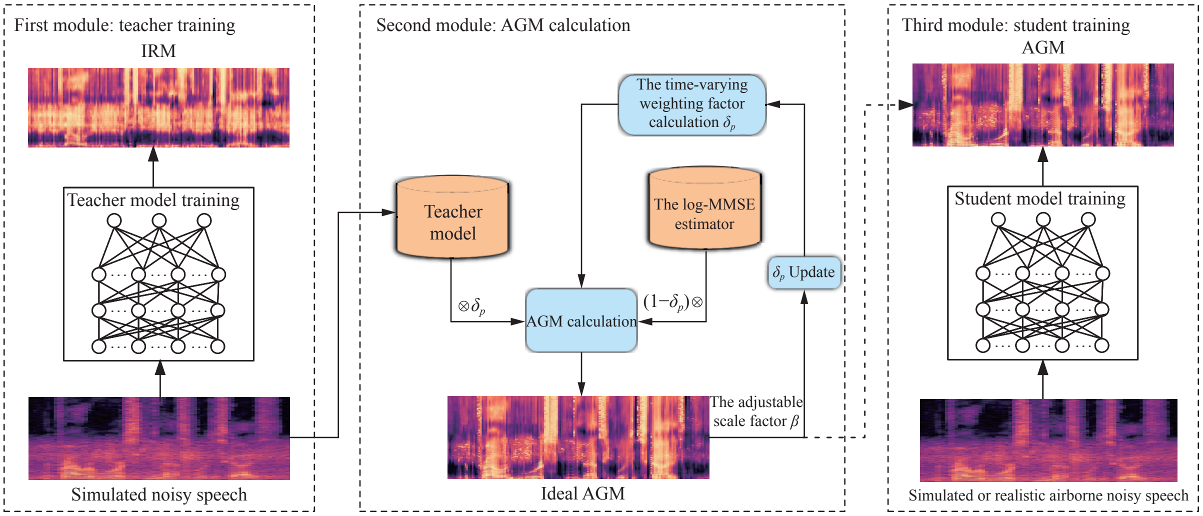

First module: teacher training
IRM
Teacher model training
Simulated noisy speech
Second module: AGM calculation
The time-varying weighting factor calculation δ_p
Teacher model
The log-MMSE estimator
δ_p Update
$\otimes\delta_p$
$(1-\delta_p)\otimes$
AGM calculation
The adjustable scale factor β
Ideal AGM
Third module: student training
AGM
Student model training
Simulated or realistic airborne noisy speech

Fig. 1. Illustration of the proposed AGM-based teacher-student training framework.

As shown in Fig.1, the purpose of the teacher model is to provide an IRM estimate, which is intended to calculate the AGM target for the student model training by combing with the conventional log-MMSE filter. Further, the second module in Fig.1, which is a well-designed online calculation procedure, aims to more flexibly and quickly utilize the advantages of those methods to update the AGM. Finally, under the guidance of the teacher model, training the student model can only require noisy speech without clean speech data. It is also the main motivation for building a student model. The details of training the whole framework and their corresponding contributions are elaborated in the following.

1. Teacher training framework

As depicted in the first module in Fig.1, the teacher model (e.g., LSTM) is exploited to learn the regressive transformation from the simulated input training data to the IRM generated by the training data pairs. In general, a deep network has the powerful modeling capacity to capture the acoustic context information in the T-F domain for obtaining an accurate IRM estimate. In [28], the estimated IRM $\hat{M}_{\mathrm{IRM}}(k,l)$ from the deep teacher model, or the spectral gain of the unsupervised method, can represent the a posteriori speech presence probability at each T-F bin. Hence, we believe the teacher model, which employs a neural network with deep architecture, has strongly complementary to the conventional log-MMSE approach for improving noise reduction performance.

In our teacher model, the input is the widely used noisy log-power spectra (LPS) feature because LPS can simulate the perceptual characteristics of the human auditory system. And the supervised fine-tuning is applied to minimize the loss function of the teacher model in (4).

2. Adaptive gain mask estimation

In this subsection, we explain the motivations for designing the proposed AGM and elaborate on how to compute it from the T-F spectral mask and the log-MMSE noise suppressor.

For the masking-based methods, the enhanced speech is usually reconstructed by utilizing the estimated mask where T-F unit values are usually mapped to the interval [0, 1]. The accuracy of the estimated mask value plays a vital role in the quality of enhancement. From [12], the log-MMSE approach can deal with quasi-stationary noise by minimizing the MSE of the log-magnitude spectra, and due to using the statistical signal processing theory, its spectral gain can provide sophisticated statistical information about the interactions between speech and background interference. No doubt, the log-MMSE suppressor, like other unsupervised methods, has difficulties in dealing with non-stationary noises. For the deep learning-based supervised algorithms, it handles non-stationary interference well by using the prior knowledge learned from the simulated training dataset, while it unavoidably brings about a high mismatch problem between the training and test dataset. Consequently, the AGM is proposed to investigate the joint effect of the spectral gain in the log-MMSE approach and the IRM in the deep learning method for speech enhancement, which is one of the AGM motivations.

Another key motivation is to more fully and autonomously concatenate the advantages of those methods for designing an accurate mask estimation which is used as a better learning target in the next section. Clearly, the weighting coefficient plays a vital role to bridge the gap between those two methods. However, in most studies, this key coefficient is arbitrarily set to be a common fixed value. The consequence of using a constant is that the process of computing the mask target treats both noise-dominant and speech-dominant frames equally which may reduce the mask's sensitivity to detect speech. Therefore, in this section, we propose a time-varying nonlinear weighting coefficient δ_P to replace the conventional fixed weighting factor.

Firstly, from (10), the gain function estimated by the log-MMSE method mainly depends on the a priori SNR $\xi(k,l)$, and $\gamma(k,l)$ is updated in (6). To improve the accuracy of estimation in adverse scenarios, the initial guess for the clean speech amplitude generated from the teacher model is utilized to estimate $\xi(k,l)$:

$$\hat{\xi}(k,l)=\frac{\left|\hat{M}_{\mathrm{IRM}}(k,l)\cdot Y(k,l)\right|^2}{\lambda_n(k,l)} \tag{13}$$

where $\hat{\xi}(k,l)$ is the estimated version of $\hat{\xi}(k,l)$. As discussed in Section II.2, $\lambda_n(k,l)$ can be obtained by using the unbiased MMSE-based noise estimator. For most traditional unsupervised speech enhancement methods, such as the constrained Wiener filtering [8]–[9], their performance can be further enhanced via using a Lagrange multiplier into the spectral gain. Motivated by this, we add the Lagrange multiplier into the traditional log-MMSE filter gain function for adjusting the tradeoff between noise reduction and speech distortion. Specifically, we insert (13) into (10) and do further calculations using a specific Lagrange multiplier μ to define a new log-MMSE filter gain function as follows:

$$G_{\text{log-MMSE}}(k,l)=\frac{\hat{\xi}(k,l)}{\mu+\hat{\xi}(k,l)}\exp\left\{\frac{1}{2}\int_{\hat{v}(k,l)}^{\infty}\frac{e^{-t}}{t}dt\right\} \tag{14}$$

$$\mu(k,l)=\mu_0-(\mathrm{SNR_{dB}})/s \tag{15}$$

$$\hat{v}(k,l)=\frac{\hat{\xi}(k,l)}{\mu+\hat{\xi}(k,l)}\gamma(k,l) \tag{16}$$

where $\mathrm{SNR_{dB}}$ denotes the noisy SNR. μ_0 and s are two empirically chosen constant which is set to 4.2 and 6.25, respectively. Next, the estimated IRM $\hat{M}_{\mathrm{IRM}}(k,l)$ from the teacher model is incorporated to define the AGM ($M_{\mathrm{AGM}}(k,l)$):

$$M_{\mathrm{AGM}}(k,l)=\delta_P\hat{M}_{\mathrm{IRM}}(k,l)+(1-\delta_P)G_{\text{log-MMSE}}(k,l) \tag{17}$$

Notably, the values of M_{AGM} can be restricted in the range of [0, 1], similarly to the IRM target, therefore M_{AGM} can be regarded as a novel ideal mask. During inference, δ_P as the time-varying weighting coefficient is derived by the average AGM at all frequency bins of the frame $l-1$, which can make a balance between $\hat{M}_{\text{IRM}}(k,l)$ and $G_{\text{log-MMSE}}(k,l)$, and it can be calculated by the following weighting function:

$$\delta_P = \frac{1}{1+\beta\times(\bar{M}_{\text{AGM}}(l-1)-1)^2} \tag{18}$$

where $\bar{M}_{\text{AGM}}(l-1)$ indicates the average AGM of the previous frame. β is an adjustable scale factor. Note that β will help to expand the range of δ_P values which actually controls the AGM estimation, and resultantly the degradation of enhancement performance in the adverse airborne environments can be controlled by selecting appropriate β values. The variation of the time-varying weighting function using different β values is shown in Fig.2. Notably, the value range of all weighting factors δ_P is limited in [0, 1] because it represents the ratio between $\hat{M}_{\text{IRM}}(k,l)$ and $G_{\text{log-MMSE}}(k,l)$.

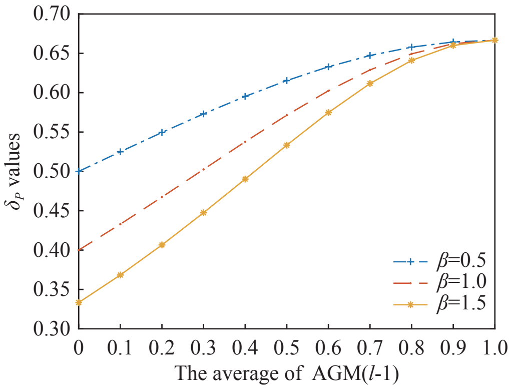


Fig. 2. The time-varying weighting function of the AGM estimation for different β values.

As shown in Fig.2, the minimum or maximum values of δ_P can be independently modified by adjusting β values, which certainly will help to accomplish an optimal tradeoff between the log-MMSE filter and IRM-based deep learning algorithm. Moreover, by using β, the updated process of δ_P is steadily varied with the historical information of AGM. Thus, we can generate a better learning target M_{AGM} for enhancing the adaptability and robustness of the student model in the unexpected acoustic scenario. After conducting tests, the final value of β is set to 1.5, because it can achieve a satisfying performance in the later experiments (cf. Section V.1).

The procedure of AGM estimation is elaborated in Algorithm 1. Clearly, the AGM is a ratio mask representation that can simultaneously describe the masking properties of the human auditory system and the statistical properties of the interactions between speech spectrum and noise. More importantly, the flexible characteristic of the time-varying weighting coefficient provides the preferred proportion between $\hat{M}_{\text{IRM}}(k,l)$ and $G_{\text{log-MMSE}}(k,l)$ at each T-F bin, which help our AGM target become a better mask expression and resultantly makes a contribution to improving enhancement performance.

Algorithm 1 The process of AGM estimation

```
Input: $Y(k,l)$: a noisy speech spectrum; $\beta$: an adjustable scale factor; $\hat{M}_{\text{IRM}}(k,l)$: the IRM estimated by the trained teacher model; $\lambda_n(k,l)$: the noise power spectrum estimation using the unbiased MMSE-based method of Section II.2;
Output: $M_{\text{AGM}}(k,l)$: the estimated AGM
Initialize: $\delta_P(k,0)=0.6$;
1: Start the following calculation process from the first T-F unit;
2: Calculating the a posteriori SNR: $\gamma(k,l)=\frac{|Y(k,l)|^2}{\lambda_n(k,l)}$;
3: Calculating the a priori SNR:
       $\hat{\xi}(k,l)=\frac{|\hat{M}_{\text{IRM}}(k,l)\cdot Y(k,l)|^2}{\lambda_n(k,l)}$
4: Calculating a spectral gain of Log-MMSE method:
       $G_{\text{log-MMSE}}(k,l)=\frac{\hat{\xi}(k,l)}{\mu+\hat{\xi}(k,l)}\exp\left\{\frac{1}{2}\int_{\hat{v}(k,l)}^{\infty}\frac{e^{-t}}{t}dt\right\}$
5: Updating the proposed time-varying weighting factor:
       $\delta_P=\frac{1}{1+\beta\times(\bar{M}_{\text{AGM}}(l-1)-1)^2}$
   where the $\beta$ is set to 1.5 according to the results in Section V.1;
6: $M_{\text{AGM}}(k,l)=\delta_P\hat{M}_{\text{IRM}}(k,l)+(1-\delta_P)G_{\text{log-MMSE}}(k,l)$;
7: Repeat through all T-F units.
```

3. Student training framework

According to the definition of the conventional mask target, it can be observed that the training work of the masking-based neural network, such as the IRM-based teacher model, demands the time-synchronized dataset with separately known clean and noise data. However, these training data pairs are extremely scarce in the airborne environment because collecting noise in the airborne environment is more difficult than in other scenarios and the existing public noise repositories have little airborne noise. To relieve this data constraint, in this section, we employ our AGM as a learning target for the student model. Specifically, the regression student model using a deep network is directly employed to estimate the frame-level AGM targets from noisy LPS features.

As illustrated in Fig.1, the dotted line presents the process of obtaining the ideal AGM (M_{AGM}). The parameters in the student model are randomly initialized, and then optimized by the stochastic gradient descent algorithm taking the MSE criterion in a mini-batch mode:

$$L_{\mathrm{SMSE}} = \sum_{k,l} \left\| \hat{M}_{\mathrm{AGM}}(k,l) - M_{\mathrm{AGM}}(k,l) \right\|_2^2 \tag{19}$$

where $\hat{M}_{\mathrm{AGM}}(k,l)$ is the estimated AGM from the student model. In the testing stage, the student model is directly exploited to decode and reconstruct the waveform.

The main contribution of the student model in our work is expected to achieve higher speech quality and better speech preservation without any constraints on the underlying clean training data. Specifically, instead of directly using the clean and noise data pairs (e.g., IRM target in (2)), the learning target AGM of the student model is calculated by the noisy speech spectrum and the well-trained teacher model. Consequently, with the guidance of the teacher model, we can replace the complex training data pairs with the simple noisy data to train the student model, which alleviates the demands for clean speech utterances and foreground noise. This is quite valuable for realistic acoustic applications. In addition, for the conventional IRM-based algorithm, its enhancement performance is usually severely degraded when there is a high mismatch between the simulated development set and the test data of the realistic airborne applications. In contrast, our student model has good robustness because its output target AGM is flexible enough to adaptively merge auxiliary guidance from the teacher model and the statistical knowledge between clean and noisy signals from the log-MMSE approach. Furthermore, owing to the inherent characteristics of the data-driven technique, the student model does not exploit any statistical assumptions about speech or noises. These merits are also of vital importance in practical application.

IV. Experimental Setup

1. Training and test set

To assess the speech enhancement performance of the proposed method under different acoustic application scenarios, we build our noise dataset, as listed in Table 1, which consists of both various airborne noises and common noises. As shown in Table 1, we provide two types of noisy speech development datasets, namely, the public training set and the airborne training set. The public training set is comprised of some widely used noises acquired from two public corpora, while the airborne training set is built by several realistic airborne noises sampled from aircraft A. All experiments in this paper adopt both the public and airborne training set as the training data except for some works in Section V.4. Notably, the key difference between those two datasets is that the airborne training set contains real-world interferences, meaning that it can be used as a noisy speech-only dataset in Section V.4. And our test set consists of the airborne noises sampled from aircraft B and other noises extracted from the Aurora corpus [33].

For all experiments in this paper, we use the Timit corpus [34] as the clean speech training and test set, which contains 6300 clean utterances spoken by 630 speakers from different dialect divisions of American English. We use 4000 clean utterances from the Timit training set to construct our training set. Then the 200 clean speech signals from the complete Timit test set are corrupted by unmatched test noises in Table 1 at five SNR levels (-7, -5, 0, 5, and 10 dB) to build our test set. Hence, the 12000 training utterances and the 1000 testing utterances are generated in total.

Table 1. Composition of the noise dataset in the experiments

Dataset	Noise source	Noise type	Noise number
Public training set	Noisex92	[35]	115
	Nonspeech	[36]	
Airborne training set	Real-time acquisition in the cockpit of aircraft A	Engine fire alarm, aircraft taxiing noise, aircraft take-off noise, aircraft landing noise, aircraft fault noise, and stall alarm.	11
Dataset	Noise source	Noise type	Noises number
Test set	Real-time acquisition in the cockpit of aircraft B	Aircraft tail noise, high-frequency metal scratching noise, propeller noise, and space noise in the aircraft cabin.	11
	Aurora	[33]	

2. Training setting

For both the training and test datasets, the sampling frequency rate is 16 kHz. The frame length for the discrete Fourier transform (DFT) is 32 ms (512 samples) with a frame shift of 16 ms (257 samples). For each frame, a Hamming window is applied. Then the 257-point LPS feature vectors are extracted to use as the inputs of all deep models.

In this paper, we attempt to provide some LSTM-based speech enhancement experiments to evaluate the performance of the proposed algorithm. Specifically, "IRM-LSTM" is the above-mentioned IRM-based method in Section II.1, which is also used as the teacher model in our experiments. We also select two well-known ratio masks (namely IBM [18] and PSM [17]) as the reference targets. Their corresponding reference supervised methods are respectively denoted as "IBM-LSTM" and "PSM-LSTM", in which a single LSTM model is employed to estimate IBM or PSM. Notably, "A

GM-TS-LSTM” denotes the proposed AGM-based method using the same LSTM architectures for both teacher and student models. For a fair comparison, we adopt the same LSTM model architecture for those methods in all experiments.

The architecture of the baseline LSTM has two hidden layers with 1024 units and a fully connected layer. Note that the input of the LSTM model in all experiments is the 257-dimensional LPS features without frame expansion. The activation function of the output layer is linear. Moreover, the Adam algorithm [37] is employed to optimize LSTM parameters in 4000 epochs, and the batch size is 500 samples. All algorithms in our work use the above-mentioned experimental settings, and all the tuning parameters in the unbiased MMSE-based noise estimation algorithm are set according to [32]. Specifically, in our work, we employ the empirical fixed value $10\log_{10}(\xi_{H_1}) = 15\ dB$ in the unbiased MMSE-based approach. All experiments were conducted with Pytorch on a workstation equipped with an Intel R CoreTMi7 CPU at 3.8 GHz.

3. Evaluation metrics

In this paper, we will report speech enhancement evaluation results based on three traditional objective measures, namely, the perceptual evaluation of speech quality (PESQ) [38], the Log-likelihood ratio (LLR) [39], and the logarithmic spectral distance (LSD) [40]. The PESQ with the range [−0.5, 4.5] is mainly used to evaluate noise reduction, and a higher PESQ illustrates better speech quality. The LLR with the range [0, 2] measures the mismatch between the formants of the clean and the enhanced signal. And the LSD is the logarithmic spectral distance measure for assessing signal similarity. In short, the LLR and the LSD can be expressed as the evaluation metric for speech distortion. For both LLR and LSD, a lower score indicates better perceptual intelligibility of the speech signal.

V. Experimental Results and Analysis

1. Comparison of the adjustable scale factor

From the aforementioned theoretical analysis in Section III.2, the adjustable scale coefficient β is quite important for the proposed AGM-based teacher-student framework. The weighting function in (18) employs the appropriate β values can create a better learning target which improves the enhancement performance of our framework. Therefore, in this section, to verify the accuracy of the selected adjustable scale coefficient β, Table 2 shows the average results of all the evaluation metrics for the AGM-based teacher-student framework with different β values. In Table 2, the adjustable scale coefficients $\beta \in [0.5, 1, 1.5]$ are examined. All results in Table 2 are presented with a 95% confidence interval. All tests are based on the same AGM-TS-LSTM architecture for speech enhancement, and other experimental settings are the same.

Table 2. The average PESQ, LSD and LLR score comparison for the proposed AGM-TS-LSTM framework with varying β values (0.5, 1, and 1.5)

PESQ (score)	SNR (dB)					
	10	5	0	−5	−7	Avg
$\beta = 0.5$	2.7066±0.07	2.3703±0.08	2.0165±0.09	1.6658±0.1	1.5392±0.11	2.0597±0.094
$\beta = 1$	2.8551±0.07	2.5139±0.07	2.1463±0.08	1.7705±0.1	1.6234±0.11	2.1818±0.09
$\beta = 1.5$	2.9029±0.078	2.5559±0.08	2.1707±0.088	1.7764±0.1	1.6224±0.11	**2.2057±0.093**
LSD (score)	SNR (dB)					
	10	5	0	−5	−7	Avg
$\beta = 0.5$	2.4998±0.189	2.6134±0.18	2.7173±0.17	2.7785±0.16	2.7842±0.16	2.6786±0.17
$\beta = 1$	2.2753±0.188	2.353±0.18	2.4267±0.17	2.4644±0.159	2.4623±0.15	2.3963±0.17
$\beta = 1.5$	2.2322±0.189	2.3069±0.179	2.3741±0.169	2.3986±0.158	2.3946±0.15	**2.3413±0.17**
LLR (score)	SNR (dB)					
	10	5	0	−5	−7	Avg
$\beta = 0.5$	0.386±0.04	0.5497±0.05	0.7559±0.06	0.985±0.068	1.0763±0.068	0.7506±0.6
$\beta = 1$	0.3661±0.04	0.5291±0.05	0.7339±0.06	0.9592±0.065	1.0472±0.065	0.7271±0.05
$\beta = 1.5$	0.332±0.04	0.4917±0.049	0.6931±0.058	0.918±0.06	1.009±0.06	**0.6889±0.05**

As depicted in Table 2, we can find that the PESQ, LSD, and LLR scores are highly related to the factor β. As a result, the speech distortion and residual noise can be balanced by choosing an appropriate β value. Specifically, compared with using other β values, the AGM-TS-LSTM, using $\beta = 1.5$, achieves the best PESQ, LSD and LLR results at each SNR. Meanwhile, for a lower β value ($\beta = 0.5$), the phenomena of speech spectral distortion and poor noise suppression are more serious. Thus, in this paper, the adjustable scale factor β is set to 1.5 which helps the whole adaptive weighting way to obtain better enhancement performance.

2. Comparison of the weighting coefficient

In this section, to evaluate the effectiveness of the proposed AGM with a time-varying nonlinear weighting coefficient δ_P in speech enhancement, we make a fair comparison of the performance of our teacher-student framework with respect to δ_P. Fig.3 shows the average PESQ, LSD, and LLR scores for the proposed teacher-student framework with different values of δ_P on the test dataset. Each value in Fig.3 is presented with a 95% confidence interval.

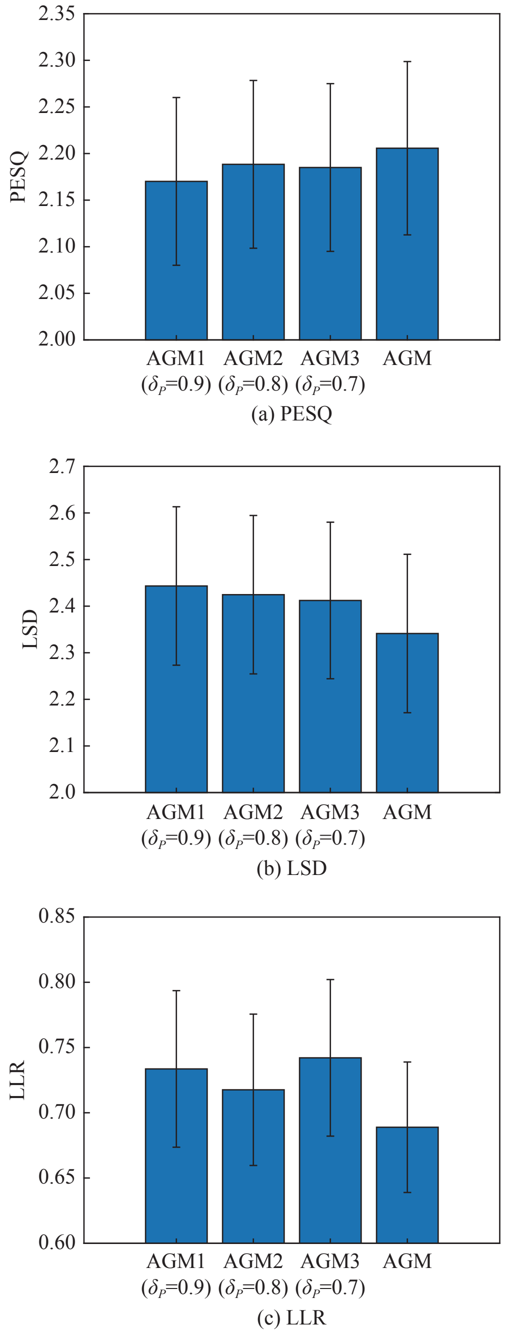
In Fig.3, we employ three typical constant values (0.9, 0.8, 0.7) of δ_P to design a learning target for the student model, denoted as AGM1, AGM2, AGM3, respectively. And "AGM" represents the proposed learning target for the student model which is calculated by using the proposed δ_P in (18). Accordingly, their corresponding enhancement framework named "AGM1-TS-LSTM", "AGM2-TS-LSTM", "AGM3-TS-LSTM", and "AGM-TS-LSTM" (proposed method), respectively. The teacher and student model of our method is the LSTM model, and they all adopt the same other parameter settings.

From Fig.3, we observed that the AGM-TS-LSTM, using δ_P in (18), gives a considerable performance improvement in terms of PESQ, LSD, and LLR. Specifically, the PESQ result of our AGM-TS-LSTM is greatly superior to other AGM-based methods, and our AGM-TS-LSTM provides the lowest LSD and LLR scores, meaning that it has the least signal distortions.

These findings indicate the teacher-student framework using a fixed δ_P value cannot bridge the gap between the merits and defects of the log-MMSE method and IRM-based deep learning algorithm. In contrast, the teacher-student framework based on the AGM target calculated by δ_P in (18) achieves the desired tradeoff between good noise reduction and high listening intelligibility, even in a low-SNR complex airborne noise environment. It demonstrates the effectiveness of the proposed adaptive weighting method.

3. Performance comparison of speech enhancement

In order to demonstrate the advantages of the proposed method under the mismatched airborne application environment and unseen social activity scenario, Table 3 provides a comparison of the average PESQ, LSD, and LLR scores on the test dataset by the LSTM-based reference algorithms at all SNR levels across 14 unseen noise types. Each value in Table 3 is provided with a 95% confidence interval. For the first block of Table 3, "b1–b4" denotes 4 types of highly non-stationary realistic airborne noise, and "car, train, restaurant, airport, exhibition, subway and street" from the Aurora database represent mismatched society noise samples. All methods adopt the same LSTM architecture, and other experimental settings are the same.



(a) PESQ

(b) LSD

(c) LLR

Fig. 3. The average PESQ, LSD and LLR score comparison for the proposed teacher-student framework with different δ_P values (0.7, 0.8, 0.9 and a time-varying δ_P in (18)).

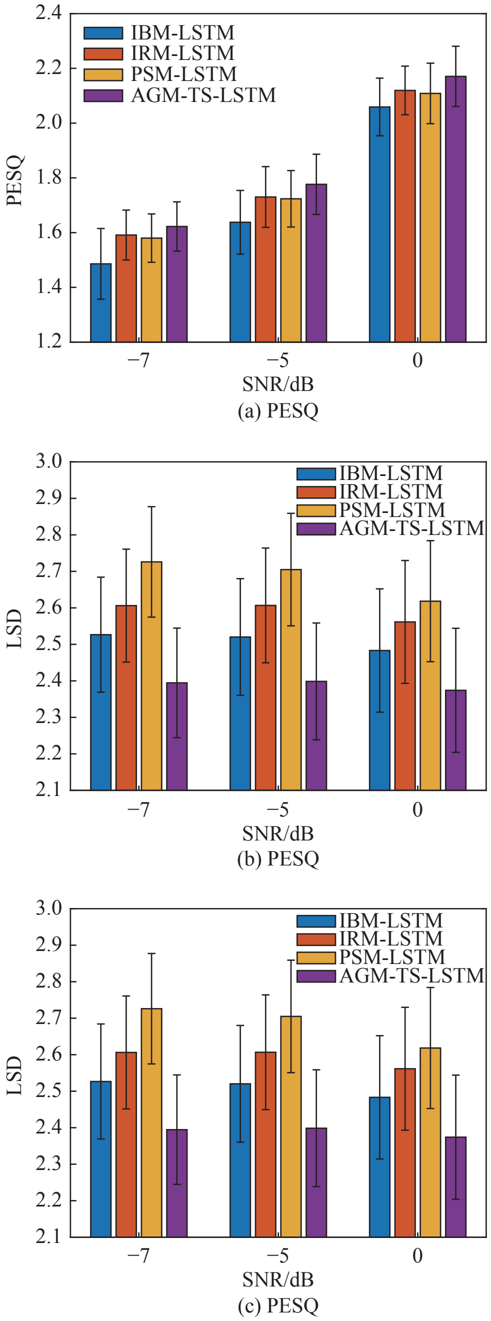
As shown in Table 3, it can be found that the enhanced speech generated by the AGM-TS-LSTM method achieves the best objective quality and intelligibility scores among those generated by the reference methods. Specifically, our AGM-TS-LSTM gives superior PESQ scores to those three methods. When compared with the PSM-LSTM and IBM-LSTM, the LSD score of our AGM-TS-LSTM gives an average decrease of 0.18 and 0.12, which account for 7.86% and 4.94% improve-

Table 3. The average performance comparisons of different LSTM-based methods on the test set across 11 unseen noise types in Table 1 at all SNR level

Noise type	PESQ			
	IBM-LSTM	PSM-LSTM	IRM-LSTM	AGM-TS-LSTM
b1	2.7432±0.13	2.7600±0.11	2.7695±0.1	2.8802±0.09
Restaurant	1.9223±0.11	1.9822±0.11	1.9777±0.1	2.0200±0.1
Car	2.0622±0.08	2.1179±0.08	2.0324±0.07	2.1368±0.085
b2	2.3160±0.08	2.2989±0.07	2.2708±0.08	2.4342±0.07
b3	2.0032±0.14	2.0911±0.12	2.1275±0.12	2.2116±0.1
Airport	2.2407±0.09	2.3501±0.09	2.3217±0.09	2.3609±0.095
Train	2.5895±0.08	2.6528±0.06	2.6965±0.07	2.6796±0.063
Exhibition	1.6335±0.11	1.6316±0.12	1.6543±0.1	1.7263±0.063
b4	1.7248±0.1	1.8451±0.1	1.7697±0.11	1.8465±0.1
Subway	1.5858±0.08	1.6215±0.1	1.6236±0.07	1.6210±0.098
Street	2.1655±0.15	2.3036±0.12	2.3047±0.11	2.3450±0.08
Average	2.0897±0.1	2.1504±0.09	2.1407±0.09	**2.2057±0.093**
Noise type	LSD			
	IBM-LSTM	PSM-LSTM	IRM-LSTM	AGM-TS-LSTM
b1	2.3600±0.29	2.2411±0.25	2.5415±0.24	2.0768±0.25
restaurant	2.2679±0.15	2.3589±0.15	2.3328±0.14	2.1009±0.15
car	2.1940±0.15	2.2853±0.16	2.4013±0.15	2.1019±0.15
b2	2.2491±0.15	2.3912±0.14	2.3390±0.15	2.1177±0.15
b3	2.1809±0.11	2.4457±0.11	2.5910±0.12	2.2634±0.11
airport	2.3075±0.2	2.3680±0.21	2.3625±0.2	2.2014±0.2
train	2.1503±0.22	2.2230±0.22	2.1883±0.21	2.0694±0.21
exhibition	2.6482±0.12	2.7275±0.12	2.7873±0.12	2.5119±0.11
b4	3.5315±0.18	3.4742±0.18	3.4839±0.18	3.2984±0.17
subway	2.9111±0.13	2.9990±0.12	3.0608±0.12	2.9208±0.12
street	2.2257±0.19	2.2653±0.2	2.2712±0.2	2.0913±0.2
Average	2.4569±0.17	2.5254±0.17	2.5781±0.16	**2.3413±0.17**
Noise type	LLR			
	IBM-LSTM	PSM-LSTM	IRM-LSTM	AGM-TS-LSTM
b1	0.8740±0.11	0.6220±0.09	0.8186±0.11	0.6091±0.09
restaurant	0.6904±0.05	0.6267±0.05	0.6504±0.05	0.5836±0.05
car	0.7013±0.05	0.6278±0.05	0.7802±0.06	0.6156±0.05
b2	0.7607±0.04	0.7657±0.05	0.8611±0.05	0.6839±0.04
b3	1.1242±0.08	0.8631±0.07	0.8809±0.07	0.8354±0.06
airport	0.5519±0.06	0.4412±0.07	0.5313±0.06	0.4507±0.05
train	0.4675±0.06	0.3035±0.05	0.3748±0.06	0.3380±0.05
exhibition	1.1267±0.07	1.0976±0.07	1.1267±0.06	1.0086±0.06
b4	1.0041±0.05	0.9285±0.06	1.0151±0.06	0.9504±0.06
subway	1.1120±0.05	1.1212±0.05	1.1845±0.05	1.1354±0.05
street	0.5019±0.05	0.3654±0.05	0.4368±0.06	0.3673±0.05
Average	0.8104±0.061	0.7057±0.06	0.7873±0.062	**0.6889±0.05**

ments, respectively. For LLR, our method achieves a 17.64% and 2.44% relative improvement over the IBM-LSTM and PSM-LSTM. Its high performance confirms that the proposed AGM target can obtain a superior noise reduction performance than IBM or PSM targets, and can avoid speech distortion as much as possible. Besides, when compared to the IRM-LSTM, the AGM-TS-LSTM yields conspicuous improvements in these three measures. It reveals the strong complementarity between the spectral gain estimated by the log-MMSE method and the predicted IRM from the LSTM-based regression algorithm. In addition, due to this strong complementarity, our method can always get a satisfactory result in terms of PESQ, LSD, and LLR, which also indicates that the proposed AGM target under the collaboration between those two methods has better performance stability than the conventional IBM, IRM, and PSM target, especially in the unexpected airborne condition.

Further, to simply and intuitively compare the enhancement performance of our approach and other reference algorithms under low SNRs, Fig.4 shows the average PESQ, LSD, and LLR scores with a 95% confidence interval at three low SNR levels (−7, −5, and 0



(a) PESQ

(b) PESQ

(c) PESQ

Fig. 4. The average PESQ, LSD and LLR scores of different LSTM-based algorithms on the test set at three low SNR levels (−7, −5, and 0 dB).

dB) across 11 unseen noise types. Obviously, the AGM-TS-LSTM method consistently achieves improvements for the three-evaluation metrics in all low SNRs scenarios. For example, when faced with background noises in the hard SNR=−5 dB case, the AGM-TS-LSTM improves significantly the PESQ from 1.63 (IBM-LSTM), 1.72 (IRM-LSTM), 1.73 (PSM-LSTM) to 1.776. The LLR is decreased from 1.06 (IBM-LSTM), 1.03 (IRM-LSTM), 0.94(PSM-LSTM) to 0.918(AGM-TS-LSTM), and the LSD score of our method is the lowest. These findings indicate our AGM target is more robust than other common targets for speech enhancement, even in heavy noise scenarios.

Overall, from the aforementioned analysis, we can find that the teacher-student training method with the AGM target performs more aggressive noise suppression and less spectral distortions than a single model with IBM, PSM or IRM targets.

4. Performance under different training data

As discussed in Section III, one significant advantage of our approach is that it can potentially be extended to any realistic noisy dataset without any specific data type (e.g., clean speech or noises). So far, in our approach, both the teacher and student training work always exploit the same training dataset. Here, in order to further explain the main motivation for our work, we analyze the implications of the realistic airborne speech data augmentation for our model training. Table 4 shows the average PESQ, LSD, and LLR comparisons of training data augmentation for our method on the test set. Each value in Table 4 is given in mean ± confidence interval (95% confidence). All tests adopt the same LSTM architecture, and other experimental settings are the same.

To keep things simple, in this work, we adopt the same LSTM model architecture. For the first block of Table 4, “AGM-TS-LSTM1” denotes the proposed AGM-TS-LSTM method in which only the public training set in Table 1 is used for the teacher and student model training. Contrary to this, in the AGM-TS-LSTM2, the student model of our AGM-TS-LSTM is trained on not only the public training set but also the airborne training set in Table 1. Notably, compared with the AGM-TS-LSTM1, AGM-TS-LSTM2 can be considered as a representative example of the data augmentation in which our method can adopt realistic airborne noisy speech to train the student model.

As depicted in Table 4, there is a particularly helpful improvement of the enhancement quality by only

Table 4. The average performance comparisons of the training data augmentation for the proposed AGM-TS-LSTM method on the test set in Table 1 at all SNRs

Method	Training data		PESQ	LSD	LLR
	Teacher model	Student model			
AGM-TS-LSTM 1	Public training set	Public training set	2.1739±0.09	2.3815±0.16	0.7044±0.05
AGM-TS-LSTM 2	Public training set	Public + airborne training set	**2.1958±0.095**	**2.3444±0.172**	**0.6829±0.05**

adding some noisy training speech in our method. Specifically, when compared with AGM-TS-LSTM1, AGM-TS-LSTM2 achieves better results in PESQ, LSD, and LLR. This high performance of the AGM-TS-LSTM2 verifies that our approach can improve speech quality and intelligibility without the restrictions on clean and noisy data pairs. This characteristic is crucial for airborne speech enhancement because the need for clean speech or noise data which is difficult to meet in real airborne scenarios can be relaxed by training the student model. Moreover, adding some noisy data sampled from real-world airborne applications also can help to alleviate the mismatch problem between the training and test data.

5. Enhanced speech spectrograms

To intuitively and simply observe the performance of our proposed AGM, in this section, we examine the resultant enhanced speech spectrograms produced by our AGM-based approach and other reference methods in Section IV.2. Fig.5 shows the magnitude spectrograms of two representative examples with the b2 (space noise in the aircraft cabin) and train noise types at different SNR levels (10 and −7 dB). We also present the spectrogram of the corresponding noisy and clean speech in Fig.5 as a reference.

From Fig.5, it can be observed that there is a lot of noise interference in the speech segment of the enhanced speech by the IBM, PSM, or IRM-based algorithm which fully reveals the drawbacks of these methods in reducing unseen noise. As expected, the speech enhanced by AGM-TS-LSTM not only has good noise removal but also retains speech segments. Specifically, as shown in the rectangular boxes on the first line in Fig.5 (e.g., b2 at 10 dB SNR case), it is clear that the AGM-TS-LSTM preserves better speech spectrum segmentation than other methods. Meanwhile, both in the typical noise condition or the extremely harsh airborne noise condition (e.g., train or b2 at −7 dB SNR level), as shown in the circled area in Fig.5, it can be seen that our method has better noise reduction performance than the other methods via using the proposed AGM estimate as a better learning target for the student model training. That is also clarified that our AGM target is more noise-robust than other standard masking targets.

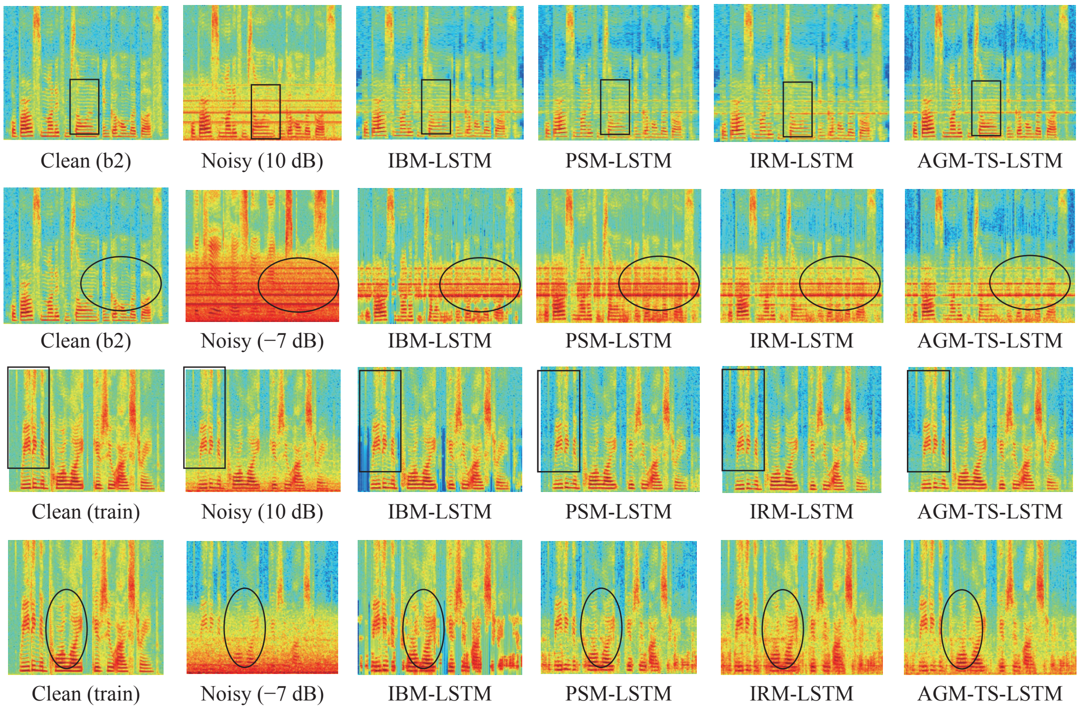


Fig. 5. Spectrogram comparison of various methods with two representative noise types (b2 and train) at the lowest and highest SNR levels (−7 and 10 dB).

In general, all objective test results in Section V show that the AGM-TS-LSTM has better enhancement performance than other reference algorithms in Airborne noise conditions or social scenarios. It demonstrates the effectiveness of the proposed AGM target based on the teacher-student framework.

VI. Conclusions

In this paper, we proposed an adaptive gain mask-based teacher-student training approach to improve the performance of LSTM-based speech enhancement. Our method combined the advantage of the log-MMSE algorithm and IRM-based deep learning techniques to design the AGM learning target for the student model. In addition, at the stage of calculating AGM, a new time-varying weighting coefficient δ_P was derived to replace the previous weighting constant. Moreover, the performance of the proposed adaptive weighting operation could be more stable by controlling the proposed adjustable scale factor β values. To test the effectiveness of our method, some experiments on the proposed method based on the LSTM model were investigated in terms of objective speech evaluation metrics. By experimental analysis, it was found that the AGM-based teacher-student training method under all noise conditions not only reduces unseen noise interference to a great extent but also achieves higher listening intelligibility.

References

[1] X. B. Cao, P. Yang, M. Alzenad, *et al.*, "Airborne communication networks: A survey," *IEEE Journal on Selected Areas in Communications*, vol.36, no.9, pp.1907–1926, 2018.

[2] S. F. Ou, P. Song, and Y. Gao, "Soft decision based Gaussian-Laplacian combination model for noisy speech enhancement," *Chinese Journal of Electronics*, vol.27, no.4, pp.827–834, 2018.

[3] S. F. Ou, P. Song, and Y. Gao, "Laplacian speech model and soft decision based MMSE estimator for noise power spectral density in speech enhancement," *Chinese Journal of Electronics*, vol.27, no.6, pp.1214–1220, 2018.

[4] W. Jiang, P. Liu, F. Wen, "Speech magnitude spectrum reconstruction from MFCCs using deep neural network," *Chinese Journal of Electronics*, vol.27, no.2, pp.393–398, 2018.

[5] T. Wang, H. Guo, B. Lyu, *et al.*, "Speech signal processing on graphs: Graph topology, graph frequency analysis and denois-ing," *Chinese Journal of Electronics*, vol.29, no.5, pp.926–936, 2020.

[6] X. Wang, Y. Guo, Q. Fu, *et al.*, "Speech enhancement using multi-channel post-filtering with modified signal presence probability in reverberant environment," *Chinese Journal of Electronics*, vol.25, no.3, pp.512–519, 2016.

[7] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.27, no.2, pp.113–120, 1979.

[8] B. Picinbono and M. Bouvet, "Constrained Wiener filtering (Corresp.)," *IEEE Transactions on Information Theory*, vol.33, no.1, pp.160–166, 1987.

[9] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol.4, no.5, pp.383–389, 1996.

[10] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Atlanta, GA, USA, pp.629–632, 1996.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.32, no.6, pp.1109–1121, 1984.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.33, no.2, pp.443–445, 1985.

[13] G. Enzner and P. Thüne, "Robust MMSE filtering for single-microphone speech enhancement," in *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, pp.4009–4013, 2017.

[14] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, pp.253–256, 2002.

[15] R. W. Li, X. Y. Sun, T. Li, *et al.*, "A multi-objective learning speech enhancement algorithm based on IRM post-processing with joint estimation of SCNN and TCNN," *Digital Signal Processing*, vol.101, article no.102731, 2020.

[16] N. Saleem and M. I. Khattak, "Multi-scale decomposition based supervised single channel deep speech enhancement," *Applied Soft Computing*, vol.95, article no.106666, 2020.

[17] S. Routray and Q. R. Mao, "Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network," *Computer Speech & Language*, vol.71, article no.101270, 2022.

[18] Z. T. Wang, X. F. Wang, X. Li, *et al.*, "Oracle performance investigation of the ideal masks," in *Proceedings of 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, pp.1–5, 2016.

[19] Q. Wang, J. Du, L. R. Dai, *et al.*, "A multiobjective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.26, no.7, pp.1185–1197, 2018.

[20] X. Y. Wang, F. Bao, and C. C. Bao, "IRM estimation based on data field of cochleagram for speech enhancement," *Speech Communication*, vol.97, pp.19–31, 2018.

[21] H. J. Yu, W. P. Zhu, and B. Champagne, "Speech enhancement using a DNN-augmented colored-noise Kalman filter," *Speech Communication*, vol.125, pp.142–151, 2020.

[22] W. L. Zhou and Z. Zhu, "A novel BNMF-DNN based speech reconstruction method for speech quality evaluation under complex environments," *International Journal of Machine Learning and Cybernetics*, vol.12, no.4, pp.959–972, 2021.

[23] G. W. Lee and H. K. Kim, "Multi-task learning U-Net for single-channel speech enhancement and mask-based voice activity detection," *Applied Sciences*, vol.10, no.9, article no.articleno.3230, 2020.

[24] N. Saleem, M. I. Khattak, M. Al-Hasan, *et al.*, "On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks," *IEEE Access*, vol.8, pp.160581–160595, 2020.

[25] D. Servan-Schreiber, A. Cleeremans, and J. L. McClelland, *Encoding Sequential Structure in Simple Recurrent Networks*. Pittsburgh: Carnegie Mellon University, 1988.

[26] L. Zhang, M. J. Wang, Q. Q. Zhang, *et al.*, "PhaseDCN: A phase-enhanced dual-path dilated convolutional network for single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.29, pp.2561–2574, 2021.

[27] Z. Y. Wang, T. Zhang, Y. Y. Shao, *et al.*, "LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement," *Applied Acoustics*, vol.172, article no.107647, 2021.

[28] Y. H. Tu, J. Du, and C. H. Lee, "Speech enhancement based on teacher-student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.27, no.12, pp.2080–2091, 2019.

[29] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol.11, no.5, pp.466–475, 2003.

[30] H. Dinkel, S. Wang, and X. N. Xu, *et al.*, "Voice activity detection in the wild: A data-driven approach using teacher-student training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.29, pp.1542–1555, 2021.

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

[32] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.20, no.4, pp.1383–1393, 2012.

[33] D. Pearce and H. G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of Sixth International Conference on Spoken Language Processing*, Beijing, China, pp.29–32, 2000.

[34] J. S. Garofolo, L. F. Lamel, W. M. Fisher, *et al.*, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM", *NIST Interagency/Internal Report (NISTIR)*, Report No.4930, 1993.

[35] A. Varga and 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 Communication*, vol.12, no.3, pp.247–251, 1993.

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

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, 2015.

[38] A. W. Rix, J. G. Beerends, M. P. Hollier, *et al.*, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs," in *Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, USA, pp.749–752, 2001.

[39] J. Beh, R. H. Baran, and H. Ko, "Dual channel based speech enhancement using novelty filter for robust speech recognition in automobile environment," *IEEE Transactions on Consumer Electronics*, vol.52, no.2, pp.583–589, 2006.

[40] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.24, no.5, pp.380–391, 1976.

HUANG Ping received the M.E. degree from Northwestern Polytechnical University in 2014. She is pursuing the Ph.D. degree at the School of Power and Energy, Northwestern Polytechnical University, Xi'an, China. Her research interests include speech enhancement and signal processing.
(Email: hp0409@mail.nwpu.edu.cn)

WU Yafeng (corresponding author) received the Ph.D. degree in electronic engineering from Northwestern Polytechnical University, Xi'an, China. He is a Professor and a Doctoral Supervisor in Northwestern Polytechnical University. His research interests include speech signal processing and vibration noise control.
(Email: yfwu@nwpu.edu.cn)