A Deep Adaptation Network for Speech Enhancement: Combining a Relativistic Discriminator With Multi-Kernel Maximum Mean Discrepancy

Jiaming Cheng, Ruiyu Liang, *Member, IEEE*, Zhenlin Liang, Li Zhao, Chengwei Huang, *Member, IEEE*, and Björn Schuller, *Fellow, IEEE*

***Abstract*—In deep-learning-based speech enhancement (SE) systems, trained models are often used to handle unseen noise types and language environments in real-life scenarios. However, since production environments differ from training conditions, mismatch problems arise that may cause a serious decrease in the performance of an SE system. In this study, a domain adaptive method combining two adaptation strategies is proposed to improve the generalization of unlabeled noisy speech. In the proposed encoder-decoder-based SE framework, a domain discriminator and a domain confusion adaptation layer are introduced to conduct adversarial training. The model has two main innovations. First, the algorithm optimizes adversarial training by introducing a relativistic discriminator that relies on relative values by applying the difference, thus avoiding possible bias and better reflecting domain differences. Second, the multi-kernel maximum mean discrepancy (MK-MMD) between domains is taken as the regularization term of the domain adversarial loss, thereby further decreasing the edge distribution distance between domains. The proposed model improves the adaptability to unseen noises by encouraging the feature encoder to generate domain-invariant features. The model was evaluated using cross-noise and cross-language-and-noise experiments, and the results show that the proposed method provides considerable improvements over the baseline without an adaptation in the perceptual evaluation of speech quality (PESQ), the short time objective intelligibility (STOI) and the frequency-weighted signal-to-noise ratio (FWSNR).**

Manuscript received May 16, 2020; revised September 14, 2020 and October 19, 2020; accepted October 19, 2020. Date of publication November 9, 2020; date of current version December 7, 2020. This work was supported in part by the National Key Research and Development Program of China under Grants 2020YFC2004002 and 2020YFC2004003, and in part by the National Natural Science Foundation of China under Grant 62001215. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Roland Badeau. *(Corresponding author: Ruiyu Liang.)*

Jiaming Cheng, Zhenlin Liang, and Li Zhao are with the School of Information Science and Engineering, Southeast University, Nanjing 210096, P.R. China (e-mail: 230198469@seu.edu.cn; zhenlinliang1@163.com; zhaoli@seu.edu.cn).

Ruiyu Liang is with the School of Information Science and Engineering, Southeast University, Nanjing 210096, P.R. China, and also with the School of Information and Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, P.R. China (e-mail: liangry@njit.edu.cn).

Chengwei Huang is with the Sugon (Nanjing) Institute of Chinese Academy of Sciences Co. Ltd., Nanjing 211106, P.R. China (e-mail: huangcwx@126.com).

Björn Schuller is with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and also with the Group on Language, Audio, and Music, Imperial College London, SW7 2AZ London, U.K (e-mail: bjoern.schuller@imperial.ac.uk).

***Index Terms*—Deep neural network, domain adaptation, maximum mean discrepancy, relativistic discriminator, speech enhancement.**

I. Introduction

SPEECH signals are easily distorted by background noises in our daily acoustic environment, thus influencing people's hearing and call quality. The distortion caused by background noises also introduce effects in many speech-related tasks (such as automatic speech recognition and speaker recognition) [1]. Therefore, the study of speech enhancement (SE), which aims to ensure the quality and intelligibility of speech while suppressing background noises, is particularly important. The research of SE has been developed for decades and has been a trending topic in the field of speech processing. It is typically challenging when the speech and noise are captured by a single microphone at the same time. This study focuses on this monaural SE task.

Classic single-channel noise suppressors are based on statistical signal processing and usually operate on the magnitude spectrogram of noisy speech, including spectral subtraction [2], Wiener filtering, the minimum mean square error (MMSE) [3] method, the minima controlled recursive averaging (MCRA) noise estimation algorithm [4] and its improved version [5]. These techniques can adapt to the noise level and perform well with quasi-stationary noises but have limitations when addressing non-stationary noise in real acoustic scenes [6]. Additionally, many unreasonable assumptions and empirical parameter settings in such algorithms limit their performances.

With the development of computer technology, supervised SE methods have been inspired by new advances in computational auditory scene analysis (CASA) [7] and machine learning, such as non-negative matrix factorization (NMF) [8], [9]. These techniques suppress noise by estimating clean speech at each time-frequency (T-F) point. In recent years, with the significant development of deep learning algorithms, data-driven SE methods have received increasing attention. In 2013, Wang extracted acoustic features from the sub-band signals in each time-frequency unit and used them as input to a deep neural network (DNN) to learn more distinguishable features [10]. Xu *et al.* [11] used a DNN based on the restricted Boltzmann

machine to reach a regression mapping, proven to be better than traditional SE methods. A SE method based on a long-short-term recurrent neural network (LSTM-RNN) was proposed in [12], which used the recursive structure between frames to capture long-term contextual information. In addition, convolutional neural networks (CNNs) [13] and generative adversarial networks (GANs) [14], [15] have also been applied to SE tasks and achieved good performances.

However, one of the key problems of data-driven SE algorithms is their generalization to untrained conditions. Due to the complexity of a real acoustic environment, the acoustic environment of an actual scene may be quite different from the acoustic environment of the training corpus. For the SE task, language type [16], noise type [17], signal-to-noise ratio (SNR) [11] and speaker identity [18] are the main issues that cause mismatched conditions. Unseen acoustic conditions can lead to an under-adaptation in supervised SE models, which severely degrades the quality of enhanced speech. A common way to solve the generalization problem is to expand the training data, that is, to train the model with as many acoustic conditions as possible. However, for complex real-world environments, it is impractical to cover an infinite amount of potential noise and language types. Further, the level of noise in the environment is constantly changing. Hence, the mismatch problem of supervised SE models always exists.

Since it is difficult to obtain clean speech corresponding to noisy speech in actual acoustic scenes, we propose to use a domain adaptive transfer learning method to solve the generalization problem. Under such circumstances, the source domain is set to have many pairs of noisy speech and clean references for training. However, only noisy speech with unseen noise types can be obtained in the target domain. Domain adaptation methods enable knowledge transfer from a labeled source domain to an unlabeled target domain via exploring domain-invariant structures that bridge different domains under substantial distributional discrepancy [19], [20]. The idea of domain adaptation has been widely used in the field of computer vision [21], [22], but it is not common in regard to the research of SE. Inspired by the research in [23], we propose a new domain adaptation framework for SE that combines two domain adaptation strategies. One of them is to build an isomorphism-inducing feature space by minimizing a distance metric of domain discrepancy to bridge the source and target domains [24], [25]. The other strategy is to use domain adversarial training to extract domain-invariant features, that is, to employ a minimax game between the feature extractor and the domain discriminator [20]. The key idea of the proposed framework is to jointly train a feature extractor and a domain discriminator. The domain discriminator tries to distinguish between the data of the source domain and that of the target domain, while the feature extractor tries to obfuscate the discriminator. During the process of domain adversarial training, we use the maximum mean discrepancy (MK-MMD) as the regularization term of the domain adversarial loss to further minimize the domain shift.

The main contributions of this paper are as follows:

- We propose a new domain adaptive framework for SE tasks that combines two transfer learning strategies. The network performs domain adaptive transfer learning through the minmax game between domain adversarial loss, domain confusion loss (MK-MMD-based loss) and regression loss, thereby improving the robustness to unseen noise.
- We explore the impact of different loss weights on domain adversarial training and verify the validity in different acoustic environments.
- We conducted separate cross-noise and cross-language-and-noise experiments, and the results have proven that the proposed algorithm can considerably improve the adaptability of the model to unseen noise and language types.

II. Related Work

In supervised SE models, untrained acoustic conditions are the main factors that cause model mismatches. However, transfer learning for SE systems is not commonly seen. Most SE models improve the generalization of the model by expanding the dataset. For example, 500 hours of training data containing 10,000 kinds of noise fragments are used in [13] to improve the performance of unseen noise types. However, the massive training data can be a burden on the computing resources, and the complexity of the model becomes correspondingly higher. Additionally, the corpus resources available for training may be scarce in actual scenarios. Sometimes, only noisy speech can be obtained, but their labels (corresponding clean speech) are unavailable. To solve such problems, it is necessary to study semi-supervised or unsupervised transfer learning methods on an incomplete dataset.

Transfer learning can build models bridging different domains and tasks by explicitly taking the cross-domain discrepancy into account [21]. In the field of speech, transfer learning has already been applied on speech recognition to adapt to unseen speakers or acoustic environments [26]. In addition, transfer learning methods have also been used to synthesize speech with different speaker identities [27]. For SE systems, Xu *et al.* [16] proposed a cross-language transfer learning method for DNN-based SE models, where the upstream network was fine-tuned to a new language, while the parameters of the downstream network were fixed in the original language. The generalization ability of SE generative adversarial networks (SEGAN) across language and noise types was investigated in [18]. However, the above algorithms are all fine-tuning methods based on the original model; they cannot adapt to the unlabeled target domain. A domain adaptive method for an SE model was first proposed by [23]. The domain adversarial training (DAT) was performed by jointly training the feature extractor and the domain discriminator and successfully adapting the model to unseen noise types. However, [23] only considered the adaptation of a single noise type in the target domain and did not explore situations with target speech signals distorted by multiple types of noises. To further improve the generalization of the SE model, we have introduced a new domain adaptive framework. The proposed framework combines two domain adaptation strategies, which are minimizing domain discrepancy distance and domain adversarial training.

For the first strategy, several methods used the maximum mean discrepancy (MMD) loss as a distance measure for domain differences. MMD computes the norm of the difference between two domain means. In [28], a two-layer neural network was trained using a denoising autoencoder for pre-training, and MMD was used as a domain confusion loss, but the shallow network lacked strong semantic representation; hence, the effect was not good. [29] added an adaptation layer to a deep convolutional neural network and used MMD to learn discriminative and domain-invariant representations based on conventional classification losses, providing closer marginal distributions between the source and target domain. MMD embedded the deep features into the reproducing kernel Hilbert spaces (RKHSs). Such kernel-embedding-based matching is relatively sensitive to the choice of kernel. Thus, Long *et al.* [19] proposed a deep adaptation network (DAN), which uses multi-kernel MMD to match deep representations across domains and multi-layer adaptation to effectively enhance the transferability of features.

For the second strategy, adversarial loss is used to minimize the domain shift, thus, learning a representation that is not able to be distinguished between domains. Tengz *et al.* [30] employed an adversarial adaptation method in a deep CNN structure to learn invariant representations by jointly optimizing for domain confusion and matching soft labels. [31] proposed the gradient reversal algorithm (Reverse Grad) to directly maximize the domain classification loss through the gradient reversal layer. The domain separation network [32] introduced the concept of a private subspace for each domain, which captures the shared representations by finding shared subspaces that are orthogonal to the private subspace. The above mentioned adversarial training methods treat domain invariance as a binary classification problem, while the adversarial discriminative domain adaptation (ADDA) [20] introduced an adversarial loss based on a generative adversarial network into the domain adversarial training. On this basis, a general adversarial adaptation framework that combines discriminative modeling, untied weight sharing, and a GAN loss was proposed.

The two domain adaptive strategies have been successfully applied to classification tasks in the fields of computer vision and speech recognition, but related research on the regression task of SE is not common [23]. In this paper, we used a SE framework based on a feature encoder-decoder structure. Our aim is to enable the feature encoder to generate domain-invariant representations through the domain adaptive method. The domain discriminator and the domain confusion adaptation layer are introduced separately, and the MK-MMD is used as the regularization term of the domain adversarial loss. Unlike the classification cross-entropy domain adversarial loss in [23], we used a GAN-based loss as the loss of the domain discriminator. To make the domain discriminator loss better reflect the distance between domains, the RSGAN-based loss [33] is introduced. In SERGAN [34], a relativistic GAN framework including a relativistic generator and a relativistic discriminator was used to perform the SE task and provided improvements. However, in our work, only the relativistic discriminator is used for domain adversarial training to do the domain adaptation of the SE methods.

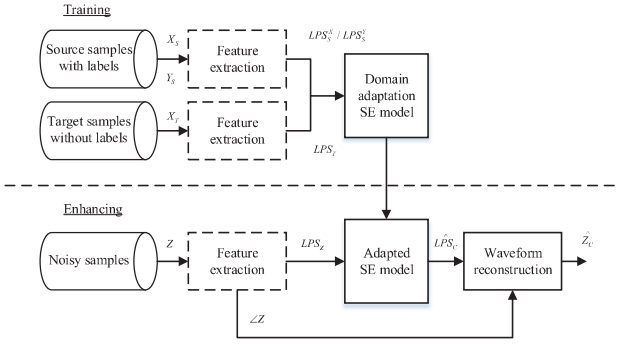


Fig. 1. Structure diagram of the proposed algorithm.

III. System Overview

This paper aims to solve the problem of domain adaptation in SE tasks. Hence, we have made the assumption that a large amount of noisy speech data and their labels (i.e., corresponding clean speech) can be obtained in the source domain. However, only the noisy speech information is available in the target domain. Our goal is to improve the performance of the SE model for the unseen environment of the target domain through a domain adaptive method. The baseline SE model follows the feature encoder-decoder structure in [23], using log-power spectra (LPS) features as the input and output of the model to perform the regression task. The structure diagram of the SE system is shown in Fig. 1.

During the training stage, the network contains two data flows. One is the noisy samples in the source domain and their corresponding labels; the other is only the noisy samples in the target domain. First, the short-time Fourier transform (STFT) is used to extract time-frequency (T-F) features of speech waveforms (X_S with label Y_S and X_T) in the two data flows and log-power spectra features (LPS_S^X with label LPS_S^Y and LPS_T) are obtained by logarithmic transformation. The proposed SE model has a feature encoder-decoder structure. Two data flows will go through the feature encoder simultaneously and share weights during the training process. Subsequently, there will be three destinations for the two data flows. Among them, only the source data flow will continue to go to the feature decoder and reconstruct the log-power spectra $\hat{LPS}_S^Y$. The mean absolute error (MAE) between $\hat{LPS}_S^Y$ and its label LPS_S^Y is calculated as the regression loss of the network. An adaptation layer is set in the second direction to compute the MK-MMD between the source domain data and the target domain data. The domain discriminator is set in the third direction to provide the relativistic adversarial loss between the domains. The method of domain adversarial training is adopted to update the parameters of the SE model. The network performs domain adaptive transfer learning through the minmax game between domain adversarial loss, domain confusion loss (MK-MMD-based loss) and regression loss. As a result, the feature encoder will learn domain-invariant features to improve the adaptability of the SE model to unlabeled target domain data.

In the enhancing stage, LPS features are extracted from the noisy signals in the target domain and implemented in the

adapted SE model to obtain the estimated clean LPS features. Considering that human listeners are not sensitive to small changes in the signal phase, the time-domain waveform of the enhanced speech is computed by an inverse Fourier transform using phase information in the noisy speech. Finally, the waveform of the entire sentence can be synthesized by an overlap-add algorithm.

IV. The Proposed Model

A. MK-MMD

Multi-kernel maximum mean discrepancy (MK-MMD) is an extension of the MMD metric. MMD is first used for the two-sample test, which accepts or rejects a null hypothesis ($\mathbb{P}=\mathbb{Q}$) based on the two samples generated from $\mathbb{P}$ and $\mathbb{Q}$. MMD uses the kernel-mapping method to embed the crucial statistical features of two distributions into the high-dimensional reproducing kernel Hilbert space (RKHS). Then, the distances between kernel mean embeddings are calculated. With MMD as the test statistic, the MMD metric between domains is equivalent to the distinguishability of the data distribution.

Let $\mathcal{H}_k$ be a reproducing kernel Hilbert space corresponding to a characteristic kernel k. Then, the kernel mean embedding of a distribution $\mathbb{P}$ in $\mathcal{H}_k$ is a unique element $\mu_k(\mathbb{P})$ such that the expectation $\mathbb{E}_{x\sim\mathbb{P}}f(x)=\langle f(x),\mu_k(\mathbb{P})\rangle_{\mathcal{H}_k}$ for all $f\in\mathcal{H}_k$. Thus, the crucial features in distribution $\mathbb{P}$ are encoded into the embedding $\mu_k(\mathbb{P})$ such that we can learn though $\mu_k(\mathbb{P})$ directly instead of $\mathbb{P}$. Then, the MMD metric between the distribution $\mathbb{P}$ and $\mathbb{Q}$ can be defined by the squared distance of the kernel mean embeddings:

$$MMD^2(\mathbb{P},\mathbb{Q}) \triangleq \|\mathbb{E}_{x^p\sim\mathbb{P}}[\phi(x^p)]-\mathbb{E}_{x^q\sim\mathbb{Q}}[\phi(x^q)]\|^2_{\mathcal{H}_k}, \quad (1)$$

where $\phi(\cdot)$ denotes a series of non-linear feature mappings in a unit sphere of a universal RKHS. Given $D_{\mathbb{P}}=\{x_i^p\}_{i=1}^m$ and $D_{\mathbb{Q}}=\{x_j^q\}_{j=1}^n$ as the sample set of the distribution $\mathbb{P}$ and $\mathbb{Q}$, an empirical estimate of MMD is:

$$MMD^2(D_{\mathbb{P}},D_{\mathbb{Q}}) \triangleq \left\| \frac{1}{m}\sum_{i=1}^{m}\phi(x_i^p)-\frac{1}{n}\sum_{j=1}^{n}\phi(x_j^q)\right\|^2_{\mathcal{H}_k}. \quad (2)$$

In the computation of MMD, the feature mapping $\phi(\cdot)$ is related to the kernel mapping $k(x^p,x^q)=\langle\phi(x^p),\phi(x^q)\rangle$. Given a characteristic kernel k in $\mathcal{H}_k$, the MMD metric can be written as:

$$\begin{aligned} MMD^2(D_{\mathbb{P}},D_{\mathbb{Q}}) \triangleq & \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(x_i^p,x_j^p) \\ & +\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(x_i^q,x_j^q) \\ & -\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i^p,x_j^q). \end{aligned} \quad (3)$$

Classical MMD is based on a single-kernel transformation. However, a single kernel is not sufficiently flexible to adequately describe various distributions. Thus, the multi-kernel MMD is proposed, which assumes that the optimal kernel function composed of multiple kernel functions can better approximate the distribution of the feature space. Multiple characteristic kernels κ can be defined as convex combinations of m kernels:

$$\kappa \triangleq \left\{ k=\sum_{u=1}^{m}\beta_u k_u : \sum_{u=1}^{m}\beta_u=1, \beta_u\geqslant 0, \forall u \right\}. \quad (4)$$

Since a Gaussian kernel can map an infinite dimensional space, a characteristic kernel based on the Gaussian function $k(x^p,x^q)=e^{\frac{\|x^p-x^q\|^2}{2\sigma^2}}$ is chosen in this paper. Multi-kernel mean embeddings can characterize distributions at different scales to match different orders of moments. In the proposed domain adaptive model, minimizing the MK-MMD metric between the source and the target domain data can decrease the edge distribution distance between the domains, so that the feature encoder tends to produce domain-invariant features. Therefore, the MK-MMD-based loss is taken as the regularization term of the domain adversarial loss.

B. RSGAN

The relativistic GAN approach presented in [33] is introduced into the domain adversarial training in this paper. The relativistic discriminator of RSGAN is used as the domain discriminator in the proposed framework.

1) Relativistic Discriminator of the RSGAN: The standard GAN [35] can be defined as:

$$\begin{aligned} L_D &= -\mathbb{E}_{x_r\sim\mathbb{P}}[\log\sigma(C(x_r))]-\mathbb{E}_{x_f\sim\mathbb{Q}}[log(1-\sigma(C(x_f)))], \\ L_G &= -\mathbb{E}_{x_f\sim\mathbb{Q}}[log(1-\sigma(C(x_f)))], \end{aligned} \quad (5)$$

where $\mathbb{P}$ and $\mathbb{Q}$ are the distributions of real data and fake data, respectively. $C(x)$ represents the non-transformed layer as $D(x)=\sigma(C(x))$, and is a sigmoid function. The principle of the standard GAN can be seen as an adversarial game between a generator and a discriminator. The generator and discriminator are viewed as a "fake producer" and a "judge". On the one hand, the "fake producer" is constantly producing fake samples in an attempt to fool the "judge". On the other hand, the "judge" continuously improves its discriminative ability to distinguish between real samples and fake samples. The two compete with each other until the "judge" is confused, and the training of the standard GAN is then completed. However, [33] argued that the key missing property of a standard GAN is that the probability of real data being real (i.e., $D(x_r)$) should decrease as the probability of fake data being real (i.e., $D(x_f)$) increases. For a standard GAN, the generator cannot affect $D(x_r)$ because the discriminator is operated independently: the real samples are not involved while training the generator, so the discriminator must remember all the attributes of real samples to guide the generator, which is a burden to the training. Therefore, [33] introduced a method of a relativistic standard GAN to make the output of the discriminator depend on the relative values of real and fake data. The discriminator is made relativistic as $D(\tilde{x})=\sigma(C(x_r)-C(x_f))$, where $\tilde{x}=(x_r,x_f)$ is sampled from real/fake data-pairs. Correspondingly, the probability that the given fake data are more realistic than randomly sampled real data is

defined as $D_{rev}(\tilde{x}) = \sigma(C(x_f) - C(x_r))$. Due to the special property of this relativistic discriminator(i.e., $1 - D_{rev}(\tilde{x}) = 1 - \sigma(C(x_f) - C(x_r)) = \sigma(C(x_r) - C(x_f)) = D(\tilde{x})$), D_{rev} do not need to be included in the loss function. The loss functions of the discriminator and the generator of the relativistic standard GAN can be written as follows:

$$L_D^{RSGAN} = -\mathbb{E}_{(x_r,x_f)\sim(\mathbb{P},\mathbb{Q})}[\log\sigma(C(x_r) - C(x_f))],$$
$$L_G^{RSGAN} = -\mathbb{E}_{(x_r,x_f)\sim(\mathbb{P},\mathbb{Q})}[\log\sigma(C(x_f) - C(x_r))]. \quad (6)$$

In this paper, the relativistic discriminator is used in the domain adversarial training. The discriminator is no longer dependent on the real/fake data-pairs, but samples from the source domain and the target domain, respectively. Theoretically, the relativistic GAN makes the discriminator only rely on the relative value by doing the difference, thus avoiding the possible bias of the discriminator and rendering the gradient more stable, which is beneficial to the domain adversarial training. Additionally, the relativistic discriminator can better reflect the distance between the source domain data and the target domain data compared with the standard GAN discriminator, which is conducive to shorten the distance between the domains.

2) Gradient Penalty in the Discriminator: Gradient penalty regularization is introduced in [36] to penalize the norm of the gradient of the critic with respect to its input, to avoid the extreme situations of the gradient (gradient vanishing and exploding) in the training process. A soft version of the constraint with a penalty on the gradient norm for random samples is used to realize the gradient penalty:

$$L_{GP}(D) = \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[\left(\|\nabla_{\hat{x}}C(\hat{x})\|_2 - 1\right)^2\right], \quad (7)$$

where $\mathbb{P}_{\hat{x}}$ is the distribution of $\hat{x} = \varepsilon x_r + (1-\varepsilon)x_f$, $x_r \sim \mathbb{P}, x_f \sim \mathbb{Q}$, and ε is sampled from a uniform distribution in [0,1]. Therefore, the loss of the discriminator after applying the gradient penalty is $L_D + \lambda_{GP}L_{GP}(D)$, where λ_{GP} is a hyper-parameter used to adjust the weight of gradient penalty. It has been observed in [33] that the application of a gradient penalty in the discriminator can stabilize the training of an RSGAN model and achieve an acceleration in convergence. Therefore, a gradient penalty is also applied to the domain adversarial training to improve the stability of the model.

C. The Proposed Algorithm

In terms of domain adaptation problems, rich source-labeled data are available, but they are not identically distributed with the target domain data. Additionally, target domain contains no labeled data points. Hence, directly adapting the original model to the target domain through fine-tuning is infeasible. Assume that there is a high-dimensional feature space that can represent the crucial features of the source and target domains, respectively. Then, such domain-invariant features can decrease the edge distribution distance between the source and target domains. In view of such a domain adaptation goal, we designed a domain adaptive framework for SE models, which mainly consists of four parts: the feature encoder ($Enc(\theta_{Enc})$) that embeds LPS features into a high-dimensional feature space, the feature decoder ($Dec(\theta_{Dec})$) that predicts clean LPS features, an adaptation layer for calculating the MK-MMD-based loss and the domain discriminator ($D(\theta_D)$) for computing the domain adversarial loss. Here, θ_{Enc}, θ_{Dec} and θ_D are parameters of the network, and the adaptation layer is only used to calculate MK-MMD metrics and will not participate in the gradient update of the network. Fig. 2 shows the overall block diagram.

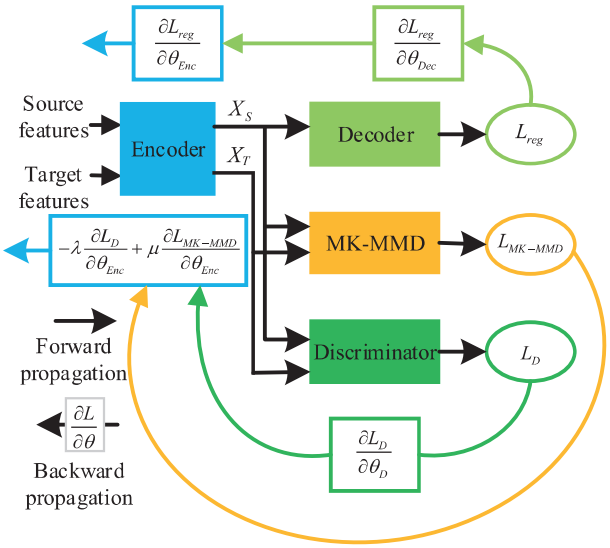


Fig. 2. The specific model structure.

The overall workflow is visualized as follows. First, the short-time Fourier transform is performed on the source domain samples (with clean labels) and target domain samples (unlabeled) to obtain the log-power spectra (LPS) features. Then, they are sent to the encoder to obtain X_S and X_T:

$$X_S = Encoder(\theta_{Enc}, LPS_S),$$
$$X_T = Encoder(\theta_{Enc}, LPS_T). \quad (8)$$

X_S and X_T are processed in the following three directions. The first way sends the labeled X_S to the decoder to reconstruct the estimated LPS features, and then the mean absolute loss (MAE) is calculated using the source domain labels:

$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N}|Dec(X_S(i)) - Y(i)|. \quad (9)$$

This loss is set to measure the performance of the source domain regression task, and minimizing it can make the model best fit the source domain dataset.

The second path computes the multi-kernel maximum mean discrepancy loss (L_{MK-MMD}) of X_S and X_T. The distribution of X_S and X_T can be made as close as possible through minimizing L_{MK-MMD}, that is, to make the encoder produce domain-invariant features that are robust to noises in different domains. $D_S = \{x_i^S\}_{i=1}^b$ and $D_T = \{x_j^T\}_{j=1}^b$ are the sample set of X_S and X_T, respectively, where b denotes the batch size. L_{MK-MMD} can be calculated by Eq. (3) with the single kernel k replaced by multiple characteristic kernels κ:

$$L_{MK-MMD} = MKMMD^2(D_S, D_T)$$
$$\triangleq \frac{1}{b^2}\sum_{i=1}^{b}\sum_{j=1}^{b}\kappa(x_i^S, x_j^S) + \frac{1}{b^2}\sum_{i=1}^{b}\sum_{j=1}^{b}\kappa(x_i^T, x_j^T)$$
$$- \frac{1}{b^2}\sum_{i=1}^{b}\sum_{j=1}^{b}\kappa(x_i^S, x_j^T), \tag{10}$$

where κ can be defined as convex combinations of h kernels as Eq. (4). The constraints on coefficients $\{\beta_u\}$ are chosen to be $1/h$. The total number of multiple characteristic kernels h is set to 19. Their parameters σ^2 are 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 5, 10, 15, 20, 25, 30, 35, 100, 1e3, 1e4, 1e5 and 1e6.

Third, X_S and X_T are sent to the domain discriminator. It can be regarded as a classifier between the source and the target domains, the last layer of which is a dense layer with one unit using a sigmoid function as the activation. The output of the discriminator with a value of 0-1 is used to compute the relativistic adversarial loss L_D:

$$L_D = -E_{x_S\sim\widetilde{S}, x_T\sim\widetilde{T}}(log\sigma(D(x_S) - D(x_T))). \tag{11}$$

Minimizing L_D actually means that the discriminator can better distinguish between samples from the source and the target domains. Here, our goal is to conduct the minimax game between the encoder and the discriminator. As described above, the discriminator is trained to make a better judgment on the samples, while the encoder is expected to extract the domain-invariant features between the source and target domains to confuse the discriminator. The weight parameters are set to balance the battle between the two. If the discriminator cannot distinguish X_S from X_T, the transfer learning from the source domain to the target domain is successful. To implement this minmax game, a gradient reversal layer (GRL) is inserted between the discriminator and the encoder. During the forward propagation, the GRL acts as an identity transformation to keep the input unchanged. During the backpropagation, however, the GRL takes the gradient from the subsequent level and changes its sign by $-\lambda$ before passing it to the encoder to form a confrontation between the encoder and the discriminator.

The parameters of the entire network are updated using the gradient descent method. The Adam algorithm [37] is used for training. The overall update rules are as follows:

$$\theta_{Enc} = \theta_{Enc} - \alpha\left(\frac{\partial L_{reg}}{\partial\theta_{Enc}} - \lambda\frac{\partial L_D}{\partial\theta_{Enc}} + \mu\frac{\partial L_{MK-MMD}}{\partial\theta_{Enc}}\right),$$
$$\theta_{Dec} = \theta_{Dec} - \alpha\frac{\partial L_{reg}}{\partial\theta_{Dec}},$$
$$\theta_D = \theta_D - \alpha\frac{\partial L_D}{\partial\theta_D}, \tag{12}$$

where α is the learning rate, and the weight parameters λ and μ are used to balance the impacts of the discriminator loss and the MK-MMD loss on the parameter updating of the encoder.

In the parameter updating process of the entire network, the encoder is expected to generate domain-invariant features through the confrontation between the encoder and the discriminator. The introduced MK-MMD loss can be regarded as the overall regularization term of the domain adversarial loss, that is, adding a constraint to the gradient update of the model, so that it can be updated in a direction that decreases the edge distribution distance between domains, thereby promoting the effect of the overall transfer learning.

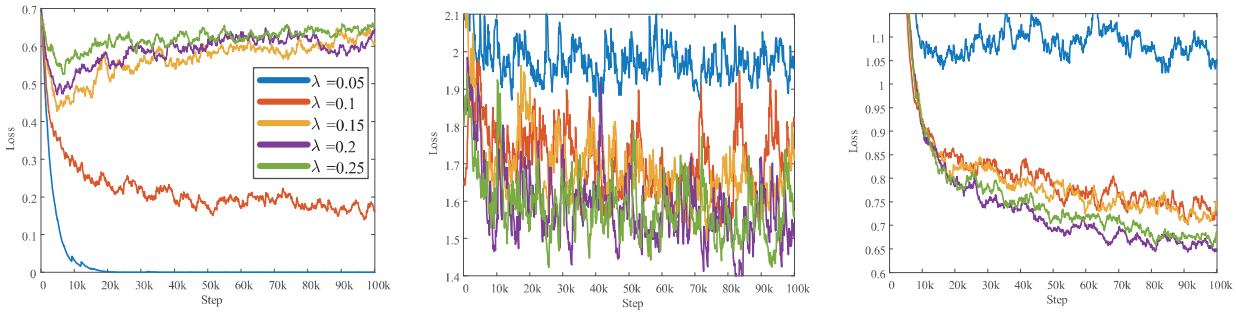
V. Experimental Results and Discussion

A. Experiment Setting

In this paper, noise type and language type are two issues that cause mismatch problems and are evaluated in the experiments. We set up two types of experiments, a cross-noise experiment and a cross-noise-and-language experiment, to test the adaptation performance of the proposed model. The cross-noise experiment is based on an English corpus in which clean utterances are selected from an excerpt of the VoiceBank-DEMAND dataset constructed by [38], containing utterances of 28 speakers from the same accent region (England) with a sampling rate of 48 kHz. For the training set, 500 utterances are randomly selected as the source domain data, and the other 500 utterances are utilized as the target domain data. The source domain data are corrupted with five matched noise types at seven SNR levels (-10 dB, -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB) to form 17,500 source domain training data. Another 500 utterances (target domain) are randomly mixed with 5 mismatched noises under the above SNR levels to form 3,500 target domain training data. The matched noise types include DestroyerEngine, FactoryFloor, HFchannel, and Pink (from NoiseX-92 [39]) as well as Wind (from the Nonspeech set [40]), while the mismatched noise types contain Speech babble (from NoiseX-92), Cry (from the Nonspeech set) and three noises in real scenes [41] (Car Riding, Crossing, and Market Place). During the training stage, the source domain data and their clean labels are used to execute the supervised training, while the target domain labels are only used to plot the loss curves but do not participate in the training. For the test set, 200 utterances selected from the VoiceBank dataset that do not cross with the training set are corrupted with five mismatched noise types at five SNR levels (-6 dB, -3 dB, 0 dB, 3 dB, and 6 dB) to evaluate the performance. For the cross-noise-and-language experiment, the source domain training data are consistent with the cross-noise experiment. However, clean utterances in the target domain are selected from the reading corpus in the Chinese Mandarin Test CD [42], which contains 60 paragraphs read by a man and a woman with a sampling rate of 44.1 kHz. In addition, 700 utterances are randomly selected from these paragraphs and divided by voice activity detection (VAD). 500 utterances are chosen for the training data in the target domain, and the rest are exploited for testing. Other settings of the experiment are consistent with the cross-noise experiment. All the datasets are available online. [1]

All the utterances and noises used in training and testing are resampled to 16 kHz. The signal frame length is 512-point (32 ms), and the frame shift is 256-point. For comparison, the baseline model is set the same as in [23]. Two bidirectional

[1][Online]. Available: https://github.com/JMCheng-SEU/audio-dataset



(a) Discriminator loss (b) MK-MMD loss (c) Target domain loss

Fig. 3. Experiments on the relativistic discriminator loss weight λ.

TABLE I
THE SETTINGS OF THE MODEL PARAMETERS

Parameters	Values
Batch size	16
Steps	100,000
Feature encoder	512 BLSTM
Feature encoder	512 BLSTM
Output layer	257 fully connected layer
Discriminator	1,024 LSTM and 1 fully connected layer
learning rate	0.0001

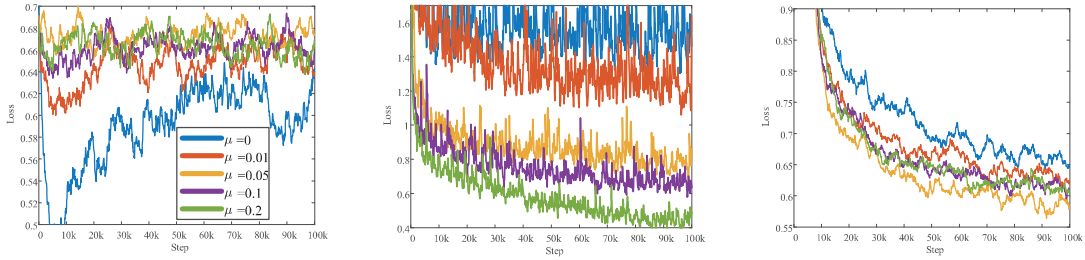
LSTM layers with 512 nodes are selected as the feature encoder and decoder, and a fully connected layer with 257 linear nodes is used as the output layer for spectrogram estimation. For the domain adaptive discriminator, unlike [23], we set a fully connected layer with one node after the unidirectional LSTM layer with 1,024 nodes and use the sigmoid function as the activation function to compute the relativistic discriminator loss instead of the categorical cross-entropy loss in [23]. The model is trained using the TensorFlow framework. The specific model parameters are shown in Table I.

B. Objective Measures for Speech Enhancement

To compare the SE performances of different methods, three objective indicators are utilized to evaluate different algorithms. The perceptual evaluation of speech quality (PESQ) [43], the short time objective intelligibility (STOI) [44], and the frequency-weighted segmental SNR (FWSNR) [45] are selected to evaluate the speech quality, speech intelligibility and the noise reduction. Among them, the PESQ is the subjective voice quality assessment indicator recommended by the ITU-T, with a score range of $-0.5\sim4.5$. The STOI measures the intelligibility of speech, and its score is between 0 and 1. The FWSNR reflects the degree of noise suppression. Compared with the global domain SNR, it is closer to the actual speech quality. For the three indicators, a higher score denotes a better result.

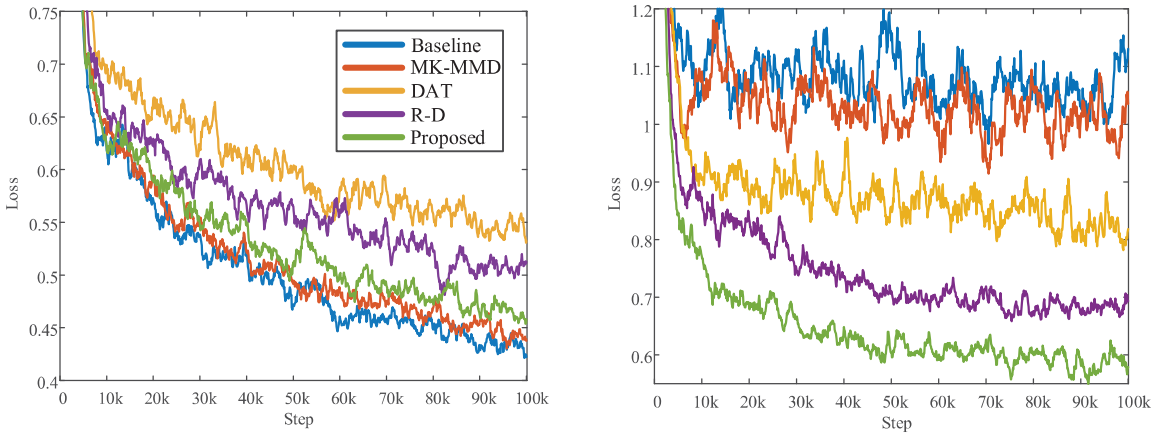
C. Determination and Analysis of Weight Parameters

In this experiment, we analyze the weight parameters λ and μ of the two domain adaptative losses in the objective function on the cross-noise experiment. First, the impact of the relativistic discriminator loss weight λ on the training is explored without introducing the multi-kernel maximum mean discrepancy loss. The value range of the weight parameter λ is $\{0.05, 0.1, 0.15, 0.2, 0.25\}$. Fig. 3 shows three types of loss curves to reflect the influence of the weight λ, that is, the MK-MMD metric loss (not involved in training, only used to show the effect) curve between encoded features from the source domain and the target domain, the discriminator loss curve and the loss curve of the reconstructed target domain samples (do not contain samples in the test set). When the weight λ is small (i.e., $\lambda = 0.05$), the loss of the relativistic discriminator will continue to decrease and converge to a minimum value after a certain number of steps. This will lead to a vanishing gradient that renders the relativistic discriminator incapable of providing reasonable guidance for the feature encoder. At the initial stage of training, the capability of the feature encoder is relatively weak, and while the discriminator is also weak, it can still distinguish between samples from the source domain and the target domain. However, the weight λ is small, so the adversarial training is insufficient. This renders the training of the relativistic discriminator to quickly saturate. It is further difficult for the subsequent network to readjust. Therefore, the target loss curve with $\lambda = 0.05$ converges to a high position. The increase of the weighting parameter λ in the interval $[0.05, 0.2]$ more obviously influences the relativistic discriminator loss. With the increase of the weight λ in this interval, the loss curve of the discriminator rises and finally converges to a higher level. This shows that strengthening the confrontation between the relativistic discriminator and the feature encoder through the weight parameter λ is conducive to the generation of domain-invariant features to a certain extent. However, when the parameter λ continues to be increased to 0.25, although the discriminator loss still rises, the loss curve of the target domain does not decrease further. This indicates that at this time, the feature encoder cannot learn more target domain knowledge, and the domain adversarial training cannot reach more transferable features by increasing the weight parameter. Therefore, $\lambda = 0.2$ is fixed as the weight of the relativistic discriminator loss to obtain a better adaptation performance. Additionally, it can be seen from the graph that the trend of the MK-MMD metric loss curve is consistent with that of the target domain, which shows that reducing the MK-MMD metric between domains can lead the adaptation training of the model to a better direction. However, it should be noted that the MK-MMD metric loss remains a high position, indicating room for optimization.



(a) Discriminator loss (b) MK-MMD loss (c) Target domain loss

Fig. 4. Experiments on the MK-MMD loss weight μ with fixed $\lambda = 0.2$.



(a) Source domain loss (b) Target domain loss

Fig. 5. Comparison of loss curves for the cross-noise experiment.

In the following stage, MK-MMD is introduced as the regularization term of the domain adversarial loss with the weight $\lambda = 0.2$. The three loss curves shown in Fig. 4 are still used to analyze the influence of the weight parameter μ. The parameter is varied in $\{0, 0.01, 0.05, 0.1, 0.15\}$. Furthermore, $\mu = 0$ indicates that MK-MMD is not introduced. It can be seen from the trend of the MK-MMD loss curve that the introduction of MK-MMD as a regularization term can effectively reduce the MK-MMD metric between the source domain and the target domain, which is not considerable when the weight parameter is small ($\mu = 0.01$). However, when the weight increases, the impact will expand. The loss of the relativistic discriminator also increases, indicating that the confrontation between domains is strengthened. The decrease of the target domain loss has proven that the introduction of the MK-MMD loss renders the model optimized towards the direction of producing domain-invariant features. Compared with the discriminator weight λ, the model is more robust to the MK-MMD weight μ. When the value of μ is higher than 0.05, although the MK-MMD loss is still reduced, the loss of the relativistic discriminator is not increased, and the target domain loss is not decreased, indicating that increasing the weight parameter of MK-MMD at this time cannot further strengthen the confrontation between domains to achieve a better adaptation effect. To balance the loss of MK-MMD and the relativistic discriminator, $\lambda = 0.2, \mu = 0.05$ are chosen as the final weight.

In theory, the fundamental purpose of domain adaptation is to obtain feature representations with domain-invariant characteristics. From this perspective, the relativistic discriminator loss improves the domain invariance of the feature encoder by executing the minmax confrontation with the feature encoder. In contrast, the MK-MMD loss reduces the cross-domain discrepancy of the feature representations in the mapping space to match different domain distributions. These findings suggest that MK-MMD reduces the domain difference actively, while the relativistic discriminator requires passive confrontation to obtain domain invariance. Therefore, the weight parameter of the relativistic discriminator is more sensitive to changes, while the weight of MK-MMD is relatively robust.

D. Cross-Noise Comparative Experiment

The same baseline architecture and weight initialization were used for all models during the experiment for a fair comparison. In addition, 200 utterances in the target domain test set are tested under 5 levels of SNR. As research on domain adaptation with unlabeled target samples for the SE task is not commonly seen, the domain adversarial training method in [23] (abbreviated as DAT) and the MK-MMD-based model (abbreviated as MK-MMD) are chosen as the unlabeled adaptation algorithms for comparison. Additionally, the fine-tuning algorithm in [18] is taken as a supervised transfer learning method for comparison. The result of the relativistic discriminator (abbreviated as R-D) is provided to verify its effect alone. Furthermore, the comparison between MMD and MK-MMD is given by indicators.

First, the loss curves of the source and target domain test set are shown in Fig. 5, which reflects the quality of the reconstructed speech from source domain samples (trained with

TABLE II
COMPARISON OF SPEECH INDICATORS FOR THE CROSS-NOISE EXPERIMENT

Indicator	Method	SNR						p[1]
		-6dB	-3dB	0dB	3dB	6dB	AVG	
PESQ	Noisy	1.421	1.689	1.867	2.034	2.224	1.847	$***$
	Baseline	1.642	1.893	2.103	2.276	2.486	2.080	$***$
	MK-MMD	1.673	1.898	2.085	2.256	2.428	2.068	$***$
	DAT	1.700	1.989	2.193	2.375	2.549	2.161	$***$
	Relativistic Discriminator	1.864	2.093	2.270	2.431	2.555	2.243	$***$
	MMD + Relativistic Discriminator	1.889	2.129	2.308	2.472	2.595	2.279	$**$
	MK-MMD + Relativistic Discriminator	1.987	2.216	2.377	2.509	2.641	2.346	/
	Fine-tune	1.846	2.140	2.350	2.528	2.713	2.315	> 0.05
STOI	Noisy	0.626	0.688	0.737	0.776	0.817	0.729	$***$
	Baseline	0.638	0.701	0.750	0.787	0.822	0.740	$***$
	MK-MMD	0.630	0.692	0.741	0.776	0.810	0.730	$***$
	DAT	0.641	0.712	0.757	0.792	0.822	0.745	$***$
	Relativistic Discriminator	0.683	0.733	0.768	0.803	0.823	0.762	$*$
	MMD + Relativistic Discriminator	0.685	0.736	0.776	0.808	0.828	0.767	> 0.05
	MK-MMD + Relativistic Discriminator	0.703	0.751	0.784	0.814	0.834	0.777	/
	Fine-tune	0.671	0.738	0.779	0.819	0.843	0.770	> 0.05
FWSNR	Noisy	1.392	2.195	3.190	4.608	5.741	3.425	$***$
	Baseline	1.990	2.936	3.919	4.950	5.812	3.921	$***$
	MK-MMD	1.894	2.814	3.772	4.758	5.601	3.768	$***$
	DAT	3.278	4.596	5.759	6.426	7.217	5.455	$***$
	Relativistic Discriminator	4.851	5.636	6.310	7.021	7.444	6.252	$*$
	MMD + Relativistic Discriminator	4.962	5.819	6.527	7.266	7.821	6.479	> 0.05
	MK-MMD + Relativistic Discriminator	5.111	5.926	6.608	7.346	7.808	6.560	/
	Fine-tune	4.683	5.862	6.599	7.447	8.126	6.543	> 0.05

[1] Note: statistic significance is shown by asterisk symbols: $*0.01 < p < 0.05$, $**0.001 < p < 0.01$, $***p < 0.001$. The symbol in the following tables have the same meaning.

TABLE III
COMPARISON OF SPEECH INDICATORS FOR THE CROSS-NOISE-AND-LANGUAGE EXPERIMENT

Indicator	Method	SNR						p
		-6dB	-3dB	0dB	3dB	6dB	AVG	
PESQ	Noisy	1.168	1.460	1.659	1.836	2.054	1.635	$***$
	Baseline	1.290	1.572	1.803	2.003	2.225	1.779	$***$
	MK-MMD	1.321	1.573	1.788	1.960	2.162	1.761	$***$
	DAT	1.328	1.577	1.827	2.011	2.233	1.795	$***$
	Relativistic Discriminator	1.400	1.656	1.874	2.040	2.230	1.840	> 0.05
	MMD + Relativistic Discriminator	1.405	1.652	1.891	2.069	2.271	1.858	> 0.05
	MK-MMD + Relativistic Discriminator	1.446	1.681	1.933	2.116	2.310	1.897	/
	Fine-tune	1.423	1.692	1.946	2.149	2.370	1.916	> 0.05
STOI	Noisy	0.652	0.721	0.785	0.832	0.880	0.774	> 0.05
	Baseline	0.650	0.720	0.783	0.827	0.867	0.770	> 0.05
	MK-MMD	0.642	0.710	0.771	0.810	0.849	0.756	$***$
	DAT	0.647	0.717	0.780	0.823	0.863	0.766	> 0.05
	Relativistic Discriminator	0.650	0.718	0.780	0.815	0.849	0.762	> 0.05
	MMD + Relativistic Discriminator	0.655	0.728	0.787	0.824	0.859	0.771	> 0.05
	MK-MMD + Relativistic Discriminator	0.660	0.731	0.788	0.827	0.860	0.773	/
	Fine-tune	0.660	0.732	0.793	0.830	0.866	0.776	> 0.05
FWSNR	Noisy	2.721	3.622	4.761	6.486	7.999	5.118	$*$
	Baseline	3.094	4.115	5.236	6.454	7.551	5.290	> 0.05
	MK-MMD	3.045	3.999	5.095	6.165	7.177	5.096	$***$
	DAT	2.935	3.870	4.974	6.063	7.096	4.988	$***$
	Relativistic Discriminator	3.504	4.417	5.425	6.179	6.991	5.303	> 0.05
	MMD + Relativistic Discriminator	3.715	4.677	5.617	6.471	7.256	5.547	> 0.05
	MK-MMD + Relativistic Discriminator	3.776	4.721	5.656	6.473	7.253	5.576	/
	Fine-tune	3.840	4.910	6.047	7.060	8.027	5.977	> 0.05

labels) and target domain samples (trained without labels). It can be seen from the loss curve of the target domain that the proposed model has the lowest loss, followed by R-D and DAT, and MK-MMD is only slightly better than the baseline model. From the comparison between domain discriminators, it can be concluded that the proposed relativistic discriminator achieves a certain improvement over the discriminator in [23]. Although MK-MMD alone has only a slight adaptation effect, its combination with the relativistic discriminator achieves an obvious improvement. This reflects that directly narrowing the distance between domain distributions does not work well. However, the MK-MMD loss can be a constraint to the domain adversarial training, making it easier to guide the model to the target domain. Notably, the source domain loss curve of each

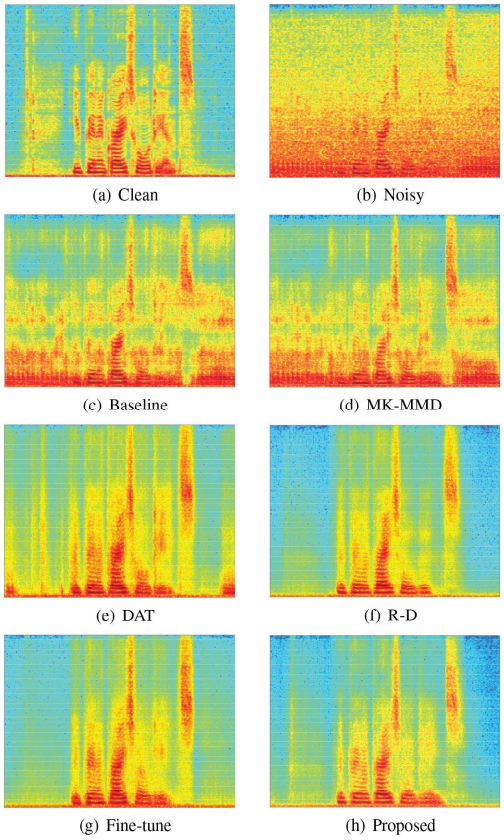


Fig. 6. Spectrograms of an English utterance (p231-166.wav) corrupted by Crossing noise (from target domain) at 0 dB SNR: (a) Clean speech, (b) Noisy speech, (c) Baseline model, (d) MK-MMD model, (e) DAT model, (f) R-D model, (g) Fine-tune model, (h) Proposed model.

model has increased compared to the baseline, which means that each transfer learning method has an adverse effect on the enhancement performance of the source domain while adapting to the target domain. Among them, DAT and MK-MMD have the most and least effect on the source domain, respectively, and the proposed combination of MK-MMD and the relativistic discriminator has an effect that lies between that of these methods. Additionally, the source domain loss of the proposed model converges to a lower position than the relativistic discriminator, which indicates that introducing MK-MMD loss as the regularization term of the domain adversarial loss can also help ensure the enhancement performance of the source domain. Essentially, the better trade-off of loss curves is brought by the constraints in objective function, and it indicates that the proposed model can improve the adaptability to the target domain while ensuring enhanced performance of the source domain.

Second, Table II shows the objective evaluation indicators of each model on the enhanced speech in the target domain. As indicated, the results contain three indicators (PESQ, STOI, and FWSNR) under five levels of SNR. Each adaptation model shows a certain improvement over the baseline, but their benefits are different. For the PESQ indicator, MK-MMD is better than the baseline only at low SNR levels (-6 dB, -3 dB), and slightly worse than the baseline at a higher SNR, indicating that reducing the MK-MMD distance between different domain distributions can improve the quality of target samples at a low SNR on this dataset. In contrast, the algorithms based on domain adversarial training, DAT and R-D, have relatively larger improvements over the baseline. In the comparison between the two, the relativistic discriminator achieves a better adaptation effect, especially at low SNR levels. The combination of MMD and the relativistic discriminator gains a further improvement, and the multi-kernel transformation is superior to the single-kernel method. Note that even compared with the fine-tuning method that uses a few target labels, the proposed model that does not require target labels still achieves a slight advantage at the average SNR and dominates at low SNR levels. This indicates that the proposed model can effectively improve the perceptual quality of speech. In terms of the speech intelligibility indicator STOI, MK-MMD performs slightly lower than the baseline, and the improvement of the DAT model is not obvious, while the relativistic discriminator achieves considerable improvement over DAT at low SNR levels. The proposed combination of MK-MMD and the relativistic discriminator still achieves the best results among all models on the average SNR, only slightly lower than the fine-tuned model under the high SNR. The comparison of the FWSNR indicator reflects the degree of noise suppression of each model. From the average results, the proposed model still ranks first among all models, which is consistent with the trends of the above two indicators. The improvement of PESQ, STOI and FWSNR over the baseline at the average SNR is 0.266, 0.037 and 2.639, respectively, which is the best result among the compared algorithms. The results of the statistical analysis show that compared with adaptation algorithms that are unlabeled, namely, DAT and MK-MMD, the improvement of all the indexes is significant ($p < 0.001$). For the supervised fine-tuning method, the proposed model can achieve slightly better results.

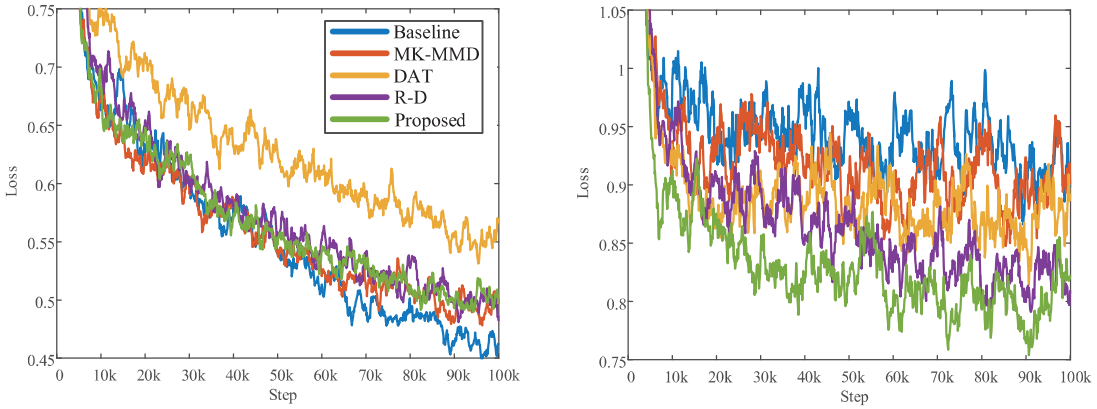
Finally, the enhanced spectrograms of the compared models are shown in Fig. 6. It can be seen that the relativistic discriminator suppresses the noise more thoroughly than the discriminator in dat. On this basis, the proposed model combining the two adaptation strategies can better retain the details of the speech.

E. Cross-Noise-and-Language Comparative Experiment

As the model may encounter unseen language types in the production environment, the adaptation performance in different language is worth testing. Additionally, the mismatched noise and language types often appear simultaneously in real scenarios, so a cross-noise-and-language experiment is performed to compare different models. The weight parameters are set the same as the experiments above to test the adaptability of the parameters to a new environment.

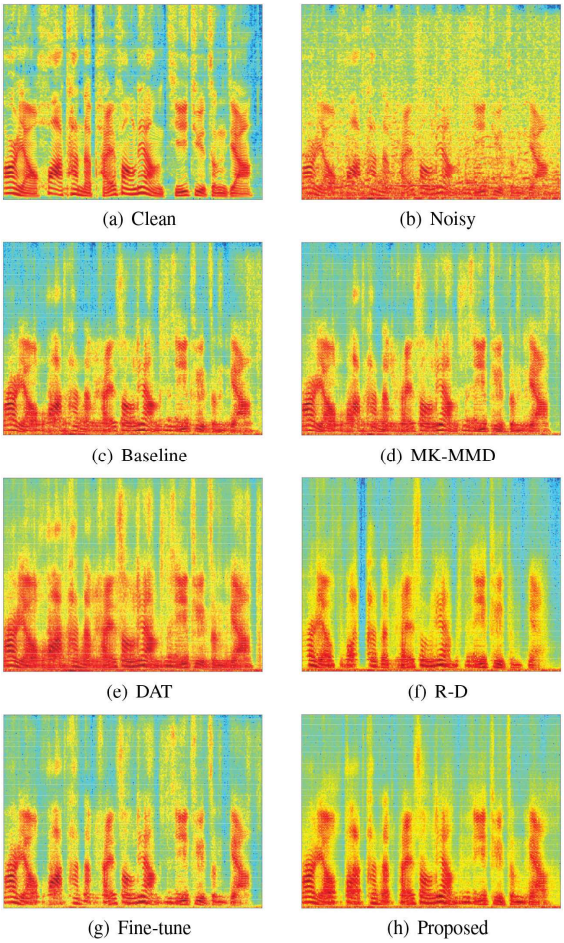
The trend of the loss curves shown in Fig. 7 is generally consistent with the results in the cross-noise experiment. However, in terms of the target domain loss, the adaptation effect of each model has obviously decreased, and the gap between models has also reduced, reflecting that the cross-noise-and-language scenario is relatively difficult.

From the perspective of specific evaluation indicators in Table III, the enhancement effect of each model is reduced compared to the cross-noise experiment. For the PESQ indicator,



(a) Source domain loss (b) Target domain loss

Fig. 7. Comparison of loss curves on the cross-noise-and-language experiment.



(a) Clean (b) Noisy

(c) Baseline (d) MK-MMD

(e) DAT (f) R-D

(g) Fine-tune (h) Proposed

Fig. 8. Spectrograms of a Chinese utterance (31-26.wav) corrupted by Speech-babble noise (from target domain) at 0dB SNR: (a) Clean speech, (b) Noisy speech, (c) Baseline model, (d) MK-MMD model, (e) DAT model, (f) R-D model, (g)Finetune model, (h) Proposed model.

the proposed model still ranks first in the unsupervised adaptation methods. However, the supervised fine-tuning approach performs better at the average SNR, only inferior to the proposed model under low SNR. In terms of the STOI indicator, all the unsupervised adaptation approaches fail to gain an effective enhancement effect at the average SNR. The results of MK-MMD, DAT and R-D are even lower than the baseline, which means they provide negative adaptation effects. The proposed method only achieves a slight improvement over the baseline under low SNR levels, indicating that the adaptation effect of speech intelligibility needs further exploration. Considering the effect of noise suppression, the proposed model achieves the highest FWSNR indexes compared with unsupervised methods but is lower that the fine-tuning method at each SNR level. In general, the proposed model can maintain a good adaptation effect that is similar to the fine-tuning method when confronting mismatched noise and language types.

Finally, the comparison of the spectrograms in Fig. 8 shows the effectiveness of the proposed algorithm again.

VI. Conclusion

In this work, we investigated the problem of noise mismatch in SE systems. We proposed a domain adaptive method for the SE task that combines two adaptation strategies, utilizing unlabeled target domain noisy speech as the guidance. A domain discriminator and a domain confusion adaptation layer are introduced, respectively, for adversarial training based on an encoder-decoder SE framework. The relativistic discriminator loss is chosen as the domain adversarial loss to better measure the differences between domains and improve the adaptation effect. Further, MK-MMD-based loss is introduced as the regularization term of the domain adversarial loss to further reduce the edge distribution distance between domains. The entire network performs domain adaptive transfer learning through the minimax game between the domain adversarial loss, MK-MMD-based loss and regression loss. As a result, the feature encoder will learn domain-invariant features to improve the adaptability of the SE model. The experimental results show that the proposed algorithm can considerably improve the adaptability of the baseline model to unseen noise and language types in the target domain. In comparison with other algorithms, the proposed algorithm achieves the best results in the unsupervised adaptation methods, and its adaptation effect is similar to the supervised fine-tuning method.

Future work includes the following aspects. First, as the improvement of speech intelligibility is not obvious, improved adaptation algorithms that aim at further improving the intelligibility of unseen noisy speech should be studied. Second, although the proposed model is effective on the speech denoising task, the applications on speech dereverberation and speaker separation are also worth exploring. Additionally, apart from the regression task in this paper, the proposed domain adaptive method is expected to be conducted in classification tasks.

REFERENCES

[1] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, 1983.

[2] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[4] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.

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

[6] Y. Xu, J. Du, and C. Lee, "Speech enhancement based on teacherc student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2080–2091, Dec. 2019.

[7] G. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, pp. 297–336, 1994.

[8] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[9] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 17–20.

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

[11] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[12] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 3709–3713.

[13] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 189–198, Jan. 2019.

[14] S. Pascual, A. Bonafonte, and J. Serr, "SEGAN: Speech enhancement generative adversarial network," in *Proc. INTERSPEECH*, 2017, pp. 3642–3646.

[15] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," 2017, *arXiv:1709.01703*.

[16] Y. Xu, J. Du, L. Dai, and C. Lee, "Cross-language transfer learning for deep neural network based speech enhancement," in *Proc. 9th Int. Symp. Chin. Spoken Lang. Process.*, 2014, pp. 336–340.

[17] D. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectr.*, vol. 54, no. 3, pp. 32–37, Mar. 2017.

[18] S. Pascual, M. Park, J. Serr, A. Bonafonte, and K. Ahn, "Language and noise transfer in speech enhancement generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5019–5023.

[19] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3071–3085, Dec. 2019.

[20] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2962–2971.

[21] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[22] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," 2009, *arXiv:0902.3430*.

[23] C. Liao, Y. Tsao, H. Lee, and H. Wang, "Noise adaptive speech enhancement using domain adversarial training," in *Proc. INTERSPEECH*, 2019, pp. 3148–3152.

[24] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[25] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2200–2207.

[26] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1450–1463, 2016.

[27] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4475–4479.

[28] M. Ghifary, W. B. Kleijn, and M. Zhang, "Domain adaptive neural networks for object recognition," D.-N. Pham and S.-B. Park, Eds. *PRICAI 2014: Trends in Artificial Intelligence. PRICAI 2014. Lecture Notes in Computer Science,*, Cham : Springer, 2014, pp. 898–904.

[29] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[30] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4068–4076.

[31] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[32] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, 2016, pp. 343–351.

[33] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," 2018, *arXiv:1807.00734*.

[34] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 106–110.

[35] I. Goodfellow *et al.*, "Generative adversarial Nets," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[36] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," 2017, *arXiv:1704.00028*.

[37] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[38] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA Held Jointly Conf. Asian Spoken Lang. Res. Eval. (O-COCOSDA/CASLRE)*, 2013, pp. 1–4.

[39] 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 Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[40] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.

[41] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification: An overview of Dcase 2017 challenge entries," in *Proc. 16th Int. Workshop Acoust. Signal Enhancement*, Tokyo, 2018, pp. 411–415.

[42] Y. Xishuang and L. Zhaoxiong, "Implementation summary of Mandarin Chinese test," Beijing, China: Commercial press, 2004.

[43] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[44] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[45] M. Jianfen, H. Yi, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, 2009.

Jiaming Cheng is currently working toward the Ph.D. degree with Southeast University, China. His research interests include speech enhancement and machine learning.

Ruiyu Liang (Member, IEEE) received the Ph.D. degree from Southeast University, China, in 2012. He is currently an Associate Professor with the Nanjing Institute of Technology, Nanjing, Jiangsu province, China. His research interests include speech signal processing and signal processing for hearing aids.

Zhenlin Liang is a Postgraduate Student with the School of Information and Communication Engineering from Southeast University, China. His research interests include deception detection and machine learning.

Li Zhao received the B.E. degree from the Nanjing University of Aeronautics and Astronautics, China, in 1982, the M.S. degree from Suzhou University, China, in 1988, and the Ph.D. degree from the Kyoto Institute of Technology, Japan, in 1998. He is currently a Professor with Southeast University, China. His research interests include speech signal processing and pattern recognition.

Chengwei Huang (Member, IEEE) received the bachelor's degree and Ph.D. degree from Southeast University, China, in 2006 and 2013, respectively. He was an Associate Professor with Soochow University from 2013 to 2014. He started a robotics company as a partner and the general Manager in 2015 focusing on natural human-computer interaction. Since 2017, he joined the Sugon (Nanjing) Institute of Chinese Academy of Sciences as CTO in big data technologies.

Björn Schuller (Fellow, IEEE) received the Diploma in 1999, the doctoral degree for his study on Automatic Speech and Emotion Recognition in 2006, and his habilitation and Adjunct Teaching Professorship in the subject area of Signal Processing and Machine Intelligence in 2012, all in electrical engineering and information technology from TUM in Munich/Germany. He is Full Professor of Artificial Intelligence, and Head of GLAM – the Group on Language, Audio & Music, Imperial College London, U.K., Full Professor and ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, Co-Founding CEO and current CSO of audEERING, and an Associate of the Swiss Center for Affective Sciences at the University of Geneva. Dr. Schuller is president emeritus of the Association for the Advancement of Affective Computing (AAAC), elected member of the IEEE Speech and Language Processing Technical Committee, , and senior member of the ACM. He coauthored five books and more than 700 publications in peer reviewed books, journals, and conference proceedings leading to more than 20 000 citations.