A Novel Adversarial Training Scheme for Deep Neural Network based Speech Enhancement

Samuele Cornell, Emanuele Principi, Stefano Squartini
Department of Information Engineering
Università Politecnica delle Marche
Ancona, Italy
s.cornell@pm.univpm.it, {e.principi,s.squartini}@univpm.it

Abstract—**In this work, we propose a novel representation-learning technique for Deep Learning-based Speech Enhancement algorithms inspired by Domain-Adversarial training. A gradient reversal layer and an additional network are employed, only at training time, to explicitly enforce a representation that is orthogonal to the additive noise in the input signal. We show that such learning scheme, which can be applied easily to most mask-based Deep Neural Network Speech Enhancement approaches, is able to improve the denoising performance when used in conjunction with scale-invariant signal-to-distortion ratio loss and allows to reach state-of-the-art performance with no computational overhead at run-time. In particular, on the commonly used VoiceBank-DEMAND benchmarking dataset, we improve signal-to-distortion ratio and signal-to-noise ratio over the non-adversarial model and CSIG, COVL and CBAK over other, state-of-the art, adversarial training techniques.**

Index Terms—**Speech Enhancement, Adversarial Training, Deep Learning, Deep Neural Networks.**

I. Introduction

Recently, monaural Speech Enhancement (SE) has seen a significant leap in performance thanks mainly to the adoption of supervised-learning techniques based on Deep Neural Networks (DNNs). In fact, DNN models, have been shown to be able to significantly improve intelligibility measures like the Short-Time Objective Intelligibility measure (STOI) [1], a feat which was not possible with non-supervised classical approaches unless very constrained situations were considered [2]. In the supervised approach, the model learns a direct mapping between noisy input features and output clean features by minimizing a loss function between the ground truth clean example and the output of the model.

A common approach is to use hand-crafted features such as Short-Time-Fourier Transform (STFT) spectra or log-spectra but, because of the high capacity of DNN models, recently, it has also been shown that it is possible to learn directly from the raw waveform by integrating the feature extraction step in the model architecture [3]–[6]. The advantage of this End-to-End approach is that the model can learn by itself the most suitable signal representation for the task at hand, possibly learning also how to exploit phase information. In fact, even if overlooked in the past, it has been shown that accurate phase reconstruction has non-trivial impact on the quality of enhanced speech [7]. While some techniques were proposed to tackle the problem of phase reconstruction from STFT of noisy signal such as the use of Phase sensitive mask (PSM) [8], or iterative phase reconstruction procedures [9], [5], accurate phase reconstruction still remains a challenging feat. On the other hand, recently, End-to-End approaches have been shown to outperform ideal STFT time-frequency masks in Source Separation [10] and regarding SE, End-to-End training was shown to be a promising direction by several works [3], [5], [6].

Aside from the choice of features, another performance critical issue regards the choice of the loss function. A widely adopted loss function is the Mean Squared Error (MSE), but more recently the Scale Invariant Signal-to-Distortion Ratio (SI-SDR) [11] has been proposed. Over the last years several other loss functions have been proposed, some of which directly borrowed from performance metrics such as STOI [6]. A comparison between MSE, SI-SDR and other monaural SE-oriented losses has been made by Kolbaek et al. [6] where it was found that SI-SDR is the most promising all-around loss function for monaural SE.

Another approach is to use the Generative Adversarial Network (GAN) framework to tackle the SE problem as it has been proposed in [3] and [12], where the loss function is derived in an adversarial way from another network which is jointly trained together with the network that is used to perform denoising. More recent works [13], [14] also based on GANs have shown further improvements by adopting a Time-Frequency masking approach and regularization techniques. However, GANs are still notoriously hard to train even if some techniques such as Spectral Normalization [15] have been able to alleviate the issue.

In this work, we propose a further technique inspired by Domain-Adversarial Training (DAT) [16] where an additional branch with a DNN preceded by a gradient reversal layer (GRL) is added to the original architecture at training time and is trained adversarially with respect to the original architecture to predict a noise-related quantity, rather than a domain label as in [16] or a noise-type label as in [17]. By competing with this additional DNN, the original architecture learns to extract a representation which is orthogonal to the noise in the input signal thus boosting denoising performance at test time. The proposed technique is applied to an End-to-End DNN to perform monaural SE, and we compare our results with other aforementioned adversarial training techniques based on

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

GANs, with an End-to-End architecture [3] and without [13], [14].

The paper is structured as follows: in Section II we briefly describe related techniques such as GANs and Domain Adversarial Training. The proposed approach is then described in detail in Section III. In Section IV, we describe the DNN architecture used for the experiments. Following, in Section V we describe the dataset, the evaluation metrics adopted and we present and discuss the results obtained. Finally, in Section VI we draw conclusions and outline possible future work.

II. Related Work

A. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [18] are powerful generative models based on a pair of neural networks that work adversarially: a generator G and a discriminator D. The generator G learns to map samples $\mathbf{z}$ from a prior distribution $\mathcal{Z}$ to samples $\mathbf{x}$ of a target distribution $\mathcal{X}$. The discriminator D, on the other hand, learns to distinguish between samples from the target distribution and samples generated by G. The two networks are jointly trained and the whole training process is formulated as a two-player minimax game between G and D as

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim\mathcal{X}}[\log D(x)] + \mathbb{E}_{z\sim\mathcal{Z}}[\log(1 - D(G(z)))],$$

where $\mathbb{E}_{x\sim\mathcal{X}}$ denotes the expectation over all the samples x coming from the distribution $\mathcal{X}$. Because as noted in [18], it often happens in practice that $\log(1 - D(G(z)))$ saturates because in early training stages G can be overpowered easily by D, practically G is usually trained to maximize $\log(D(G(z)))$ instead. This GAN formulation goes by the name of non-saturating GAN [19].

The aforementioned SE-oriented GAN approaches [3], [12], [14] and [13] are actually conditional GANs as the generator takes as input a noisy signal and tries to approximate the clean signal. Moreover, different, more robust formulations such as Least-Square GANs [20] (LSGAN) coupled with a regularization loss to help stabilize training are employed. In detail, in [3], LSGAN is coupled with L_1 regularizing loss which is added to the generator loss. In [12] the Pix2Pix [21] approach (originally developed for images) is adopted, while in [14] and [13] the standard non-saturating GAN loss is employed, but a MSE regularizing loss is added to the generator as it is observed that without such regularizing loss, G learns easily how to fool D without producing acoustically convincing results.

B. Domain-Adversarial Training

Domain Adversarial Training [16] is a domain adaptation technique that tackles the mismatch between a training and a target domain by enforcing a model to learn features that are invariant to the change of domains. This is achieved by embedding the domain adaptation process into the training procedure by adding, to the original architecture, a branch with a gradient reversal layer followed by a domain classifier. These two added components are only used at training time and then dropped at test-time, so there is no computational overhead at run-time. During training, both the original network task and this newly added domain classification task are jointly optimized. The gradient reversal layer encourages that the feature extraction stage of the original architecture works adversarially to the added domain classifier, by extracting features that are domain-invariant and thus maximize the loss of the domain classification task. Liao et al. [17] have shown that such technique can be also employed in monaural Speech Enhancement to increase the robustness of the model to unseen noise types and thus improve algorithm generalization.

III. Proposed Approach

The proposed training scheme is closely related to DAT and is illustrated in Figure 1. As we are performing classical SE without dereverberation, we assume that the input noisy signal $\mathbf{y} = [y_1, y_2, \ldots, y_M]$ is given by

$$\mathbf{y} = \mathbf{x} + \mathbf{v}, \tag{1}$$

where $\mathbf{x}$ is the clean speech signal and $\mathbf{v}$ an additive unknown noise signal. Moreover, we denote as $\mathbb{Y} = \{\mathbf{y}_i\}_1^K$ the input noisy examples and with $\mathbb{X} = \{\mathbf{x}_i\}_1^K$ the corresponding clean examples in the training dataset, where K is the total number of examples. As shown in Figure 1 we have chosen to adopt a masking approach motivated by the encouraging results from [13] and [14] but without making any assumption on the input transformation which thus can be learnt in an End-to-End fashion. This transformation can be either STFT spectra or log-spectra or another arbitrary (linear or non-linear) signal transformation, hand-crafted or learnt as in [4], [10]. In this framework, the transformed representation for the noisy signal $\mathbf{Y}$ is thus obtained via an analysis transformation $\mathcal{A}(\mathbf{y})$ while the estimated clean transformed representation $\tilde{\mathbf{X}}$ is obtained through a neural-network estimated mask $\mathbf{\Omega}_c$ via element-wise multiplication (Hadamard product) as

$$\tilde{\mathbf{X}} = \mathbf{Y} \odot \mathbf{\Omega}_c. \tag{2}$$

Finally, a synthesis transformation $\mathcal{S}(\tilde{\mathbf{X}})$ is responsible for obtaining the estimated time-domain signal $\tilde{\mathbf{x}}$ from the masked transformed representation.

We assume that the network used to estimate the mask $\mathbf{\Omega}_c$ can be decomposed into two parts: an encoder part $\mathcal{E}(\mathbf{Y};\theta_e)$ with θ_e parameters, where higher-level features are extracted from the input signal transform $\mathbf{Y}$ and a decoding/masking part $\mathcal{M}(\mathcal{E}(\mathbf{Y};\theta_e);\theta_m)$ with θ_m where these features are aggregated in order to produce the speech-related mask $\mathbf{\Omega}_c$.

In our experiments, we have chosen to train this network End-to-End by using learnable analysis $\mathcal{A}(\mathbf{y};\theta_a)$ and synthesis $\mathcal{S}(\tilde{\mathbf{X}};\theta_s)$ blocks as in [10] and choosing as loss function SI-SDR between reconstructed (post-synthesis) estimated clean signal and target clean signal

$$\mathcal{L}_{main}(\tilde{\mathbf{x}}, \mathbf{x}) = -10\, log\left(\frac{\|\alpha\mathbf{x}\|^2}{\|\alpha\mathbf{x} - \tilde{\mathbf{x}}\|^2}\right), \tag{3}$$

where $\alpha = \tilde{\mathbf{x}}^T \mathbf{x} / \|\mathbf{x}\|^2$ is a re-scaling factor used to enforce scale-invariance.

The core of the proposed approach consists in the fact that, at training time, an additional branch with a gradient reversal layer and an additional network is added between the encoding and decoding part of the network. Contrary to DAT where the additional network is used to classify the domain of the input examples and differently from [17] where, instead, is used to classify the noise type, here, this additional network, is used to predict a noise-related quantity such as the noise Ideal Binary Mask (IBM) or Ideal Ratio Mask (IRM) in the transformed domain. In fact, it is not the scope of this work to develop another domain adaptation technique for SE but to devise a training scheme that is able to boost the algorithm performance at training and test time without any assumption on a potential domain mismatch.

This added network $\mathcal{D}(\mathcal{E}(\mathbf{Y});\theta_d)$ takes in input encoder features and estimates a mask $\tilde{\mathbf{M}}_n$ for the noise, trying to discriminate, for each bin, the probability of noise presence. This discriminator network is thus trained via a cross-entropy loss between the estimated mask $\tilde{\mathbf{M}}_n$ and the target noise IRM or IBM

$$\mathcal{L}_{adv}(y, \mathbf{M}_n) = \frac{1}{N}\sum_{i}^{N} \mathbf{M}_n log(\mathcal{D}(\mathcal{E}(\mathcal{A}(\mathbf{y})))), \qquad (4)$$

where N is the total number of bins in the transformed domain and the target mask $\mathbf{M}_n$ is a function of $\mathbf{y}$ and $\mathbf{x}$ as, for example, IRM can be obtained as

$$\mathbf{M}_n = \frac{\mathcal{A}(\mathbf{y}-\mathbf{x})}{\mathcal{A}(\mathbf{x}) + \mathcal{A}(\mathbf{y}-\mathbf{x})}, \qquad (5)$$

and IBM by simply applying an hard decision threshold for each IRM bin.

The total loss is then a linear combination of the main and adversarial losses weighted by an hyperparameter β:

$$\mathcal{L}(\mathbf{y}, \mathbf{x}) = \beta\mathcal{L}_{main} + (1-\beta)\mathcal{L}_{adv}. \qquad (6)$$

The addition of the GRL ensures that the parameters of the encoder $\mathcal{E}(\mathbf{Y};\theta_e)$ and the discriminator $\mathcal{D}(\mathcal{E}(\mathbf{X});\theta_d)$ are updated adversarially during stochastic gradient descent at each iteration with learning rate l_r:

$$\theta_e \leftarrow \theta_e - l_r\left(\beta\frac{\partial\mathcal{L}_{main}}{\partial\theta_e} - (1-\beta)\frac{\partial\mathcal{L}_{adv}}{\partial\theta_e}\right), \qquad (7)$$

$$\theta_d \leftarrow \theta_d - l_r(1-\beta)\frac{\partial\mathcal{L}_{adv}}{\partial\theta_d}, \qquad (8)$$

$$\theta_m \leftarrow \theta_m - l_r\beta\frac{\partial\mathcal{L}_{main}}{\partial\theta_m}. \qquad (9)$$

Analysis $\mathcal{A}(\mathbf{y};\theta_a)$ and synthesis $\mathcal{S}(\tilde{\mathbf{X}};\theta_s)$ transforms are also updated in the same fashion. In this way, the encoder $\mathcal{E}(\mathbf{Y};\theta_e)$ and the analysis transform $\mathcal{A}(\mathbf{y};\theta_a)$ are encouraged to find a representation for the input signal which is orthogonal to the additive noise in order to maximize the loss of the adversarial network. The β parameter is crucial in this sense as it is responsible for the trade-off between minimization of SI-SDR and orthogonality to the noise in the feature space: SI-SDR tries to minimize the distance between the rescaled target signal and the estimated signal in time domain while the adversarial loss forces the estimated signal to be orthogonal to the noise in the encoder feature space.

Looking specifically at the update rules in Equations 7 and 8 a comparison can be made with GANs. As in GANs, in fact, the encoder $\mathcal{E}$ and discriminator $\mathcal{D}$ compete against each other. In fact, as noted in [17], DAT can be seen as a different minimax approximation beside the alternative minimization procedure used in GANs which actually leads to same formulation as a non-saturating GAN [19] with conditional input.

However, a major difference is that here the discriminator is not trying to discriminate to which distribution (clean or enhanced) the input encoder features belongs to, but instead tries to infer the noise for each bin of the input signal representation in order to produce a mask, thus leading to the following minimax formulation with value function $V(\mathcal{D}, \mathcal{E})$:

$$\min_{\mathcal{D}}\max_{\mathcal{E}} V(\mathcal{D}, \mathcal{E}) = \mathbb{E}_{\mathbf{y},\mathbf{x}\sim p_{data}(\mathbf{y},\mathbf{x})}\mathcal{L}_{adv}(\mathbf{y}, \mathbf{x}), \qquad (10)$$

where for $\mathcal{L}_{adv}$ we have explicitly considered that the target noise mask $\mathbf{M}_n$ is a function of the target clean signal $\mathbf{x}$ and noisy input $\mathbf{y}$ and we have ignored the contribution of $\mathcal{L}_{main}$ which can be seen as an additional regularization loss for the encoder $\mathcal{E}$.

IV. Network Architecture

We decided to adopt the recently proposed Dual-Path Recurrent Neural Network (DPRNN) [22] as the core component in all the different configurations we explored in our experiments. This recently developed neural architecture adopts a divide-and-conquer strategy which was shown to be very effective in Source Separation, surpassing the performance of previous techniques by a large margin while, at the same time, halving the number of parameters with respect to previous models.

As in [4], [10] we used for the analysis $\mathcal{A}$ and synthesis $\mathcal{S}$ blocks respectively a 1D convolutional layer and a 1D transposed-convolutional layer with learnable kernels. Both layers have 64 channels and a kernel size of 16 with stride 8. We settled for these hyper-parameters after some preliminary experiments. Interestingly we found that further reducing the kernel size had negative impact on the performance. This is in stark contrast to what was found for Source Separation [22] where better performance was observed with smaller kernel sizes. One possibility is that small kernel sizes work well for clean two speaker Source Separation because of the highly structured nature of human speech but fail to have enough discriminative capability when less structured input is encountered.

In the analysis block, the convolutional layer was followed by a Rectified Linear Unit (ReLU) non-linear activation to guarantee a non-negative output. In fact, IRM and IBM computation for the adversarial network targets in Equation 5 require non-negative values for the transformed representations. Encoder $\mathcal{E}$, decoder $\mathcal{M}$ and adversarial network $\mathcal{D}$ are all composed of one DPRNN block [22] each of which has two

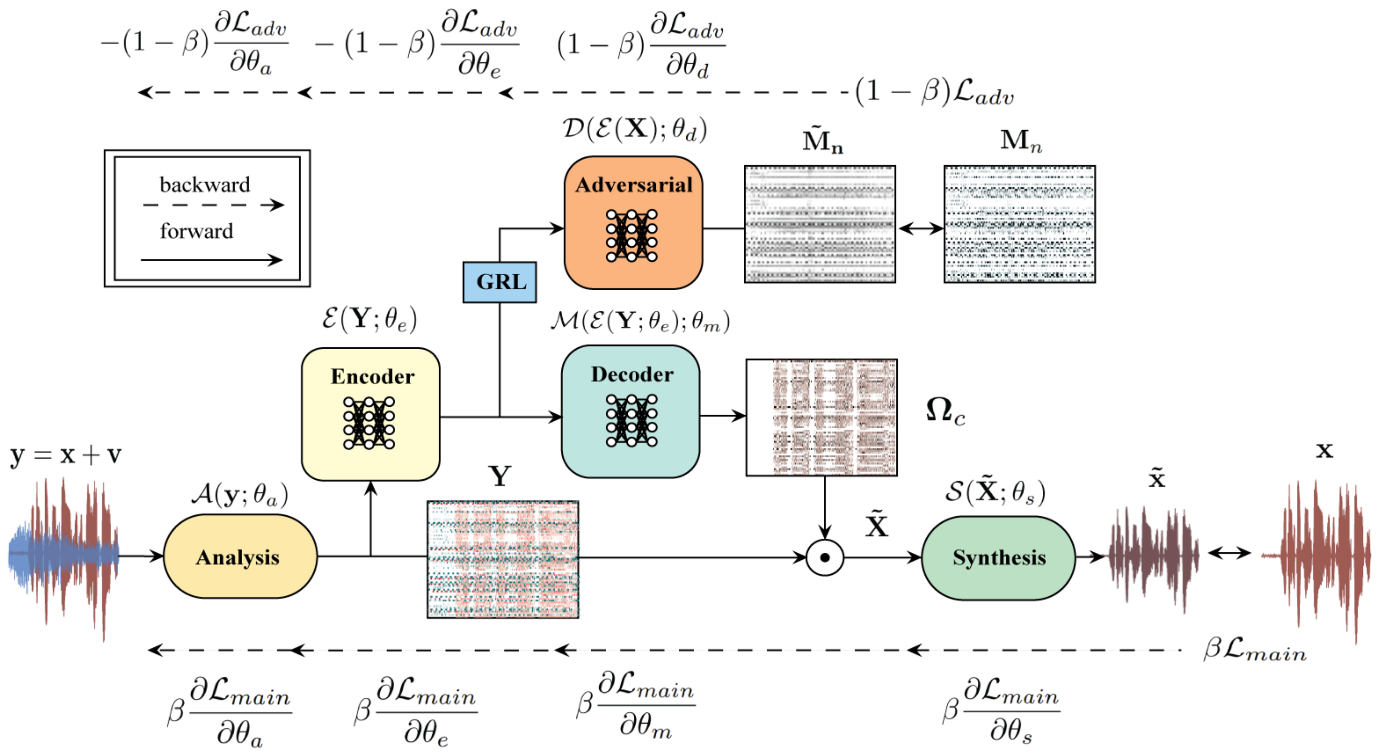
$-(1-\beta)\frac{\partial \mathcal{L}_{adv}}{\partial \theta_a}$
$-(1-\beta)\frac{\partial \mathcal{L}_{adv}}{\partial \theta_e}$
$(1-\beta)\frac{\partial \mathcal{L}_{adv}}{\partial \theta_d}$
$(1-\beta)\mathcal{L}_{adv}$
$\mathcal{D}(\mathcal{E}(\mathbf{X});\theta_d)$
$\tilde{\mathbf{M}}_\mathbf{n}$
$\mathbf{M}_n$
backward
forward
Adversarial
GRL
$\mathcal{E}(\mathbf{Y};\theta_e)$
$\mathcal{M}(\mathcal{E}(\mathbf{Y};\theta_e);\theta_m)$
Encoder
Decoder
$\mathbf{\Omega}_c$
$\mathbf{y}=\mathbf{x}+\mathbf{v}$
$\mathcal{A}(\mathbf{y};\theta_a)$
$\mathbf{Y}$
$\mathcal{S}(\tilde{\mathbf{X}};\theta_s)$
$\tilde{\mathbf{x}}$
$\mathbf{x}$
$\tilde{\mathbf{X}}$
Analysis
Synthesis
$\beta\mathcal{L}_{main}$
$\beta\frac{\partial \mathcal{L}_{main}}{\partial \theta_a}$
$\beta\frac{\partial \mathcal{L}_{main}}{\partial \theta_e}$
$\beta\frac{\partial \mathcal{L}_{main}}{\partial \theta_m}$
$\beta\frac{\partial \mathcal{L}_{main}}{\partial \theta_s}$

Fig. 1: Illustration of the proposed approach. It consists of an encoder, a decoder and an adversarial neural network plus an analysis and synthesis transform blocks. The adversarial network learns to predict a noise-related mask by minimizing $\mathcal{L}_{adv}$. A GRL ensures that the encoder and analysis blocks work adversarially versus the adversarial network by maximizing $\mathcal{L}_{adv}$. In this way the analysis and encoder blocks are encouraged to learn a noise-orthogonal representation.

bi-directional Long-Short-Term Memory (LSTM) networks with 128 neurons. The encoder $\mathcal{E}$ has an additional channel-wise normalization layer at its input followed by a point-wise (unitary kernel size) convolutional layer as in [10] with 64 channels.

The decoder $\mathcal{M}$ and adversarial network $\mathcal{D}$ have instead an additional 64 channel point-wise convolutional layer followed by a Parametric ReLU (PReLU) [23] non-linearity and an optional activation to produce the mask.

We use the sigmoid activation for both the decoder and the adversarial network, as it gave the best results in our preliminary experiments. Thus the decoder tries to estimate a pseudo IRM mask for the clean transformed representation which minimizes the SI-SDR in time domain between denoised and target signal.

The network is then trained till convergence via early-stopping on 4 second long segments with a batch-size of 1. RAdam [24] is used for optimization with an initial learning rate of $1e^{-3}$. We also apply gradient clipping with maximum L_2 norm of 5 to improve training stability.

V. Results and Discussion

A. Dataset

For our experiments, we used the VoiceBank-DEMAND [25] which has become a de-facto standard dataset to compare monaural Speech Enhancement algorithms. It consists of a total of 24792 utterances of 30 different speakers selected from VoiceBank corpus [26] and 10 artificially added noises some of which selected from the DEMAND dataset [27]. This dataset is already split into a training and test set with no overlap between speakers and noise types and with noises added at different signal-to-noise ratios: for the training set 15 dB, 10 dB, 5 dB and 0 dB and for the test set 17.5 dB, 12.5 dB, 7.5 dB and 2.5 dB. This makes this dataset particularly suitable for assessing an algorithm generalization capability. As the original sampling rate for the data is 48 kHz we downsampled to 16 kHz in order to be comparable with previous works [3], [13], [14] which also report performance figures on this dataset at 16 kHz. Moreover, in our experiments, the training set was further split into a training and a validation set using a 90/10 ratio. The validation set was then used for hyper-parameter tuning and early-stopping.

B. Evaluation Metrics

In this study, we use several objective measures to assess the performance of the proposed method. Alongside SI-SDR defined in Equation 3 and used as the main loss $\mathcal{L}_{main}$ in our architecture, Signal-to-Distortion Ratio (SDR), Signal-to-Artifacts Ratio (SAR) and Signal-to-Noise Ratio (SNR) as defined in [28] are reported. The popular BSS Eval toolkit [29] is used to compute these metrics.

To compare directly with other aforementioned adversarial training methods [3], [14] and [13], we also compute the wideband ITU-T P.862.2 Perceptual Evaluation of Speech Quality (PESQ) [30] (from -0.5 to 4.5) a metric originally developed for voiced telecommunication evaluation, Short-Time Objective Intelligibility (STOI) [1] and Hu and Loizou [31] composite measures: CSIG, CBAK and COVL (all from 1 to 5) which are objective measures that approximate SIG, BAK and OVL subjective Mean Opinion Scores (MOS). More in detail, CSIG quantifies the signal distortion when the listener is attending to the speech signal, CBAK instead quantifies the intrusiveness of the noise when the listener is attending to the noise and COVL the overall quality of the signal.

C. Ablation Study

To assess the validity of the proposed approach we have performed an ablation study by comparing different configurations of the same architecture. In particular, we considered two additional possible training schemes beside the proposed approach (Adv):

- Non-Adversarial (Non-Adv) where the adversarial branch is removed and the analysis, encoder, decoder and synthesis blocks are trained only via $\mathcal{L}_{main}$.
- Cooperative (Coop) where the GRL in the adversarial branch is removed and the discriminator is given the task to estimate the IRM or IBM of the clean signal instead of the one for the noise. In this configuration, differently from Equation 7, the gradients with respect to the two losses have the same sign at the encoder and analysis blocks.

For both Adv and Coop, two different masks were considered as a target: IRM and IBM. As a reference, we also computed the objective metrics outlined in Section V-B for the un-enhanced noisy input and for oracle STFT-based IBM and IRM with 25 ms window and 10 ms stride.

In our preliminary experiments, we also explored a third configuration, where the adversarial branch is interposed between the analysis and encoder block and thus only the analysis transform, besides the discriminator, is affected by $\mathcal{L}_{adv}$. However, we found out that, because of limited capacity of the learnable analysis transform, as expected, this leads to unstable training with discriminator easily overpowering the analysis block even when the discriminator capacity was severely reduced.

For all the training schemes considered, we performed a random-search based hyper-parameter tuning. In the following, we thus report only the best model measured in terms of COVL for each scheme unless stated differently. As pointed out in Section III a particularly critical hyper-parameter for the proposed approach is β which controls the trade-off between adversarial and SI-SDR loss. In Figure 2, we have reported CSIG versus CBAK for different values of β and for Adv-IRM training scheme. It can be seen that there is an optimum range of values with a good trade-off between CSIG and CBAK. The best value, in this case, is 0.8 for which the highest COVL is obtained as it offers the best trade-off. In general, we found that for values higher than 0.9 and lower than 0.6 there was a significant drop in performance.

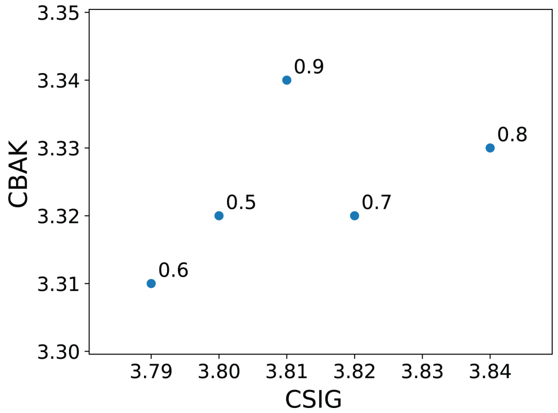

Fig. 2: CBAK versus CSIG for different values of β for the Adv-IRM proposed learning scheme.

Method	SI-SDR	SDR	SNR	SAR
Noisy	8.44	8.53	8.53	∞
Non-Adv	19.26	20.59	25.47	22.49
Adv-IBM	19.38	21.1	27.67	22.50
Adv-IRM	**19.51**	**21.52**	**28.18**	**22.57**
Coop-IBM	19.13	20.34	24.98	22.56
Coop-IRM	19.16	20.31	25.1	22.54
Oracle-IBM	19.7	20.97	30.49	21.64
Oracle-IRM	19.24	20.27	24.35	22.53

TABLE I: Ablation study results in terms of SI-SDR and BSS Eval [29] metrics. Bold-fonts indicate best performance (except for oracle).

Method	CSIG	CBAK	COVL	PESQ	STOI
Noisy	3.33	2.44	2.62	1.97	0.91
Non-Adv	3.82	3.33	3.19	2.57	**0.94**
Adv-IBM	3.82	**3.36**	**3.21**	**2.62**	**0.94**
Adv-IRM	**3.84**	3.33	**3.21**	2.59	**0.94**
Coop-IBM	3.79	3.31	3.15	2.52	**0.94**
Coop-IRM	3.80	3.31	3.17	2.54	**0.94**
Oracle-IBM	3.33	3.57	3.22	3.19	0.97
Oracle-IRM	4.94	4	4.45	3.66	0.98

TABLE II: Ablation study results in terms of composite metrics, PESQ and STOI. Bold-fonts indicate best performance (except for oracle).

In Table I we report the results in terms of SI-SDR, SDR, SNR and SAR. As it can be seen, even the baseline Non-Adv configuration attains values close to oracle STFT masks. In particular, the fact that, for all models, SAR is higher with respect to the oracle masks should indicate that the analysis and synthesis blocks learn a transformation which is able to also capture phase information leading ultimately to less artifacts in the reconstructed signal. Secondly we see that the Coop learning scheme is not able to bring benefits but instead results in a slight performance loss. The overall best figures are obtained with the proposed techniques (Adv-IBM, Adv-IRM) which are even able to surpass oracle masking for what regards SDR and SAR and even SNR but only with respect to Oracle-IRM. In particular, the one where the discriminator is used to estimate the IRM (Adv-IRM) perform best as far as these metrics are concerned.

In Table II we compare the different training schemes in terms of objective composite metrics, PESQ and STOI. In contrast with Table I here it can be seen that, in general, the different schemes are far from oracle STFT performance and especially Oracle-IRM. The only exception is for CSIG, in fact, all models achieve better CSIG than Oracle-IBM which is not able to improve over Noisy as the IBM masking approach introduces significant distortion. This result is in accordance with the SAR values and further shows that the End-to-End approach adopted leads to less artifacts in the estimated clean signal. Also in accordance with SI-SDR and BSS eval metrics, the Coop learning schemes lead to a slight overall drop in performance compared with the baseline Non-Adv scheme.

Both the proposed approaches are able to improve over the Non-Adv baseline but in different ways: Adv-IBM attains the highest CBAK but is not able to improve CSIG while, on the contrary, Adv-IRM improves CSIG but not CBAK. This seems to suggest that Adv-IBM is able to denoise more but distorts less while Adv-IRM has the exact opposite behaviour. This is in contrast with the SNR figures from Table I where the highest SNR is achieved by Adv-IRM but is in accordance with the SAR figures as the similar SAR values for Non-Adv and Adv-IBM actually lead to identical CSIG. This suggests that SNR and CBAK are poorly correlated, while SAR and CSIG are more strongly correlated. The proposed approaches are also able to bring a slight improvement in terms of PESQ, but no significant improvement has been observed for STOI. Figure 3 shows LogMel spectrograms of respectively oracle clean signal $\mathbf{x}$ (Figure 3a), noisy signal $\mathbf{y}$ (Figure 3b) and estimated clean signal $\tilde{\mathbf{x}}$ for Non-Adv (Figure 3c) and Adv-IRM (Figure 3d). It is observed that, in accordance with the metrics in Table I and II, the proposed approach leads to higher SNR in the enhanced signal.

A statistical significance test was conducted to assess if the values obtained for composite metrics, PESQ and STOI in Table II are of enough statistical significance. We resorted to the non-parametric Wilcoxon Signed-Rank Test [32] as the data distribution for the performance metrics was found to be highly-non Gaussian and because we are comparing different algorithms on the same test data. We adopted the default assumption for the test: the null hypothesis consists in assuming same distribution for the two methods we are comparing. The significance value was set to 0.01. In Table III we report the p-values obtained with such test by comparing the proposed approaches (Adv-IBM, Adv-IRM) with the baseline Non-Adv. It can be seen that the null hypothesis can be rejected for all metrics but STOI, CBAK for Adv-IRM and STOI, CSIG for Adv-IBM.

	p-value				
Method	CSIG	CBAK	COVL	PESQ	STOI
Adv-IBM	0.23	$2.5e^{-17}$	$1e^{-3}$	$6e^{-14}$	0.51
Adv-IRM	$5e^{-3}$	0.43	$2e^{-3}$	$7e^{-4}$	0.67

TABLE III: Wilcoxon statistical significance test results between the proposed adversarial methods and non-adversarial baseline method.

To further understand how the proposed approach is able to improve the performance over the plain SI-SDR only approach (Non-Adv) we can analyze the output of the encoder block. In fact, the rationale behind the proposed method is that the encoder block learns to extract a representation more robust with respect to noise by competing against the adversarial network. In order to test such hypothesis we have computed the mean $\mathcal{L}_2$ norm for the test set data and for the output of the encoder. In particular, we computed such value when only the oracle clean signal $\mathbf{x}$ is fed to the architecture $\mathcal{L}_2^{clean} = \|\mathcal{E}(\mathbf{X})\|_2$ and when only the noise $\mathbf{v}$ is fed to the architecture $\mathcal{L}_2^{noise} = \|\mathcal{E}(\mathbf{V})\|_2$. In Table IV we report such values for all the different training schemes. The fact that the ratio between $\mathcal{L}_2^{clean}$ and $\mathcal{L}_2^{noise}$ is higher for the proposed method suggests that the encoder is able to better reject noise when the proposed adversarial scheme is employed.

Method	$\mathcal{L}_2^{clean}$	$\mathcal{L}_2^{noise}$	$\mathcal{L}_2^{clean}$ / $\mathcal{L}_2^{noise}$
Non-Adv	0.092	0.113	0.81
Adv-IBM	0.088	0.092	0.95
Adv-IRM	0.087	0.091	0.95
Coop-IBM	0.075	0.09	0.83
Coop-IRM	0.078	0.093	0.84

TABLE IV: Values for the mean $\mathcal{L}_2$ norm on test data for the output of the encoder when the model is fed the oracle clean signal ($\mathcal{L}_2^{clean}$) and the oracle noise signal ($\mathcal{L}_2^{clean}$).

In Figure 4, as a further example, we show the output of the encoder when the network is fed the noisy input $\mathbf{y}$. For visualization purposes we have taken the logarithm of the absolute value of the feature maps. In the top Figure 4a is visualized the output of the encoder trained with proposed Adv-IRM learning scheme while in the bottom Figure 4b with baseline Non-Adv method. The speech has a distinctive pattern and appears as vertical stripes that encompass every channel. This suggests that every feature map is activated by speech. On the contrary, it can be seen that in both models noise triggers overall more sparse activations which shows that the encoder and analysis blocks have learnt how to extract relevant speech-related features. It is also observed that, for the proposed approach (Adv-IRM), the speech patterns are more visible which indicate that the encoder learns a more robust representation against input noise as noise leads to weaker activations in the adversarially trained model.

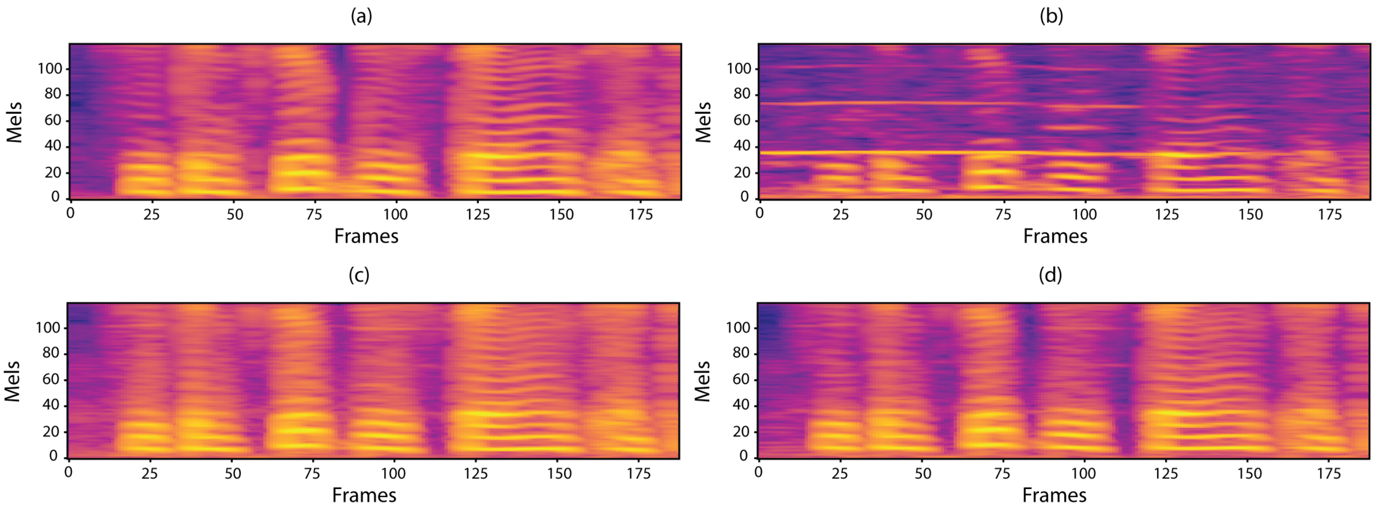

Fig. 3: LogMel spectrograms of oracle clean signal (a), noisy signal (b), estimated clean signal with Non-Adv model (c) and estimated clean signal with proposed Adv-IRM model (d).

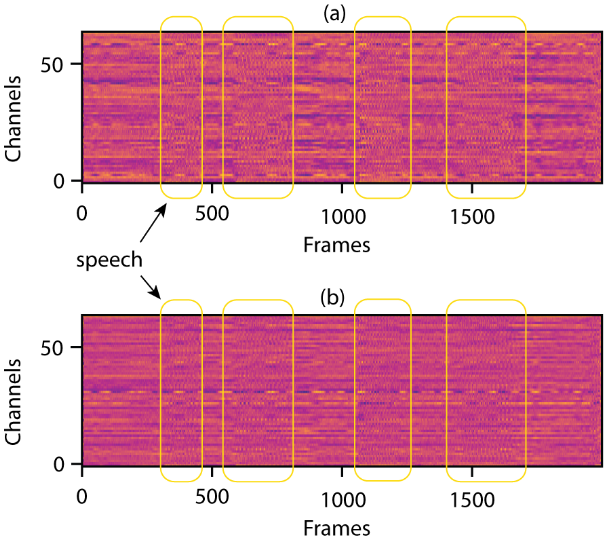

Fig. 4: (a): encoder output for Adv-IRM. (b): encoder output for Non-Adv. Speech regions have been highlighted.

D. Comparison with other Methods

In Table V we compare the results obtained in the previous Section with other state-of-the-art algorithms based on adversarial training for which composite metrics, PESQ and STOI are available on VoiceBank-DEMAND. It can be seen that both proposed learning schemes (Adv-IBM, Adv-IRM) are able to outperform the other approaches based on GANs with a smaller model at runtime. In fact, the second best algorithm [14] has about 850K parameters at runtime while both the proposed techniques 441K. In particular, it can be observed that, among the other proposed algorithms the two TF-masking approaches significantly outperform SEGAN [3] whose relatively low CSIG value suggests it tends to introduces more artifacts. We argue that such a result is due to the fact that SEGAN adopts an End-to-End regression approach, which could lead to severe non-linear distortions. On the contrary, we adopted the learnable transformation framework firstly proposed in [4] for Source Separation, where a masking approach is used together with learnable low-capacity transformations. Due to the low capacity of analysis and synthesis blocks strong non-linear distortions are avoided.

Method	CSIG	CBAK	COVL	PESQ	STOI
Noisy	3.33	2.44	2.62	1.97	0.91
Adv-IBM	3.82	**3.36***	3.21*	**2.62***	**0.94**
Adv-IRM	**3.84***	3.33	**3.21***	2.59*	**0.94**
SEGAN [3]	3.48	2.94	2.8	2.16	0.93
Soni et al. [14]	3.8	3.12	3.14	2.53	0.93
Shah et al. [13]	3.55	2.95	2.92	2.34	0.93

TABLE V: Comparison between different adversarial training methods on VoiceBank-DEMAND. For the proposed approach, statistical significant results over the non-adversarial model are denoted with an asterisk.

VI. Conclusions

In this study, we propose a novel training scheme for Speech Enhancement Deep Learning-based algorithms based on adversarial training. At training time an additional branch with a gradient reversal layer followed by an additional neural network is added to the original architecture after the feature extraction stage. This additional network is given the task to predict a noise-related IBM or IRM mask while the gradient reversal layer ensures that such network works adversarially with respect to the feature extraction stage layers.

In this way, the feature extraction stage learns to extract more discriminative features which are more robust against noise. This proposed learning scheme was applied to a End-to-End analysis/masking/synthesis deep neural network architecture where, instead of using fixed Time-Frequency transformations, the transformation and its inverse is learnt and a

DNN-predicted mask for the speech is applied on such learnt representation. We performed an extensive ablation study to assess the validity of the proposed training scheme using the VoiceBank-DEMAND dataset and comparing various training schemes with several objective measures. Results show that the proposed approach is able to improve several objective metrics over the baseline non-adversarial training scheme.

Finally, the proposed training scheme was also compared with other SE algorithms based on adversarial training for which results on VoiceBank-DEMAND are available, and it was shown to be able to outperform previous methods with significantly less parameters. Possible future work could explore how to make this framework less sensitive to the β hyper-parameter, which has currently a significant impact on the performance, and apply the proposed training scheme to different application scenarios as a pre-processing step for Automatic Speech Recognition.

References

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

[2] G. Kim and P. C. Loizou, "Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1581–1596, 2011.

[3] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[4] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 696–700.

[5] J. Kim, M. El-Kharmy, and J. Lee, "End-to-end multi-task denoising for joint sdr and pesq optimization," *arXiv preprint arXiv:1901.09146*, 2019.

[6] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *arXiv preprint arXiv:1909.01019*, 2019.

[7] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

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

[9] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," *arXiv preprint arXiv:1903.03107*, 2019.

[10] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[11] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done?" *arXiv preprint arXiv:1811.02508*, 2018.

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

[13] N. Shah, H. A. Patil, and M. H. Soni, "Time-frequency mask-based speech enhancement using convolutional generative adversarial network," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1246–1251.

[14] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5039–5043.

[15] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," in *Domain Adaptation in Computer Vision Applications*. Springer, 2017, pp. 189–209.

[17] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," *arXiv preprint arXiv:1807.07501*, 2018.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[19] W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow, "Many paths to equilibrium: Gans do not need to decrease a divergence at every step," *arXiv preprint arXiv:1710.08446*, 2017.

[20] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[22] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," *arXiv preprint arXiv:1910.06379*, 2019.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[24] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.

[25] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[26] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–4.

[27] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. ASA, 2013, p. 035081.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[29] C. Févotte, R. Gribonval, and E. Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.

[30] P.862.2, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," *ITU-T Std. P.862.2*, 2007.

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

[32] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*. Springer, 1992, pp. 196–202.