INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Improved Speech Enhancement using a Time-Domain GAN with Mask Learning

Ju Lin[1], *Sufeng Niu*[2], *Adriaan J. van Wijngaarden*[3],
Jerome L. McClendon[1], *Melissa C. Smith*[1] *and Kuang-Ching Wang*[1]

[1] Clemson University, Clemson, SC
[2] LinkedIn Inc., Mountain View, CA
[3] Nokia Bell Labs, Nokia, Murray Hill, NJ

{jul, sniu, smithmc, kwang}@clemson.edu,
adriaan.de_lind_van_wijngaarden@nokia-bell-labs.com

Abstract

Speech enhancement is an essential component in robust automatic speech recognition (ASR) systems. Most speech enhancement methods are nowadays based on neural networks that use feature-mapping or mask-learning. This paper proposes a novel speech enhancement method that integrates time-domain feature mapping and mask learning into a unified framework using a Generative Adversarial Network (GAN). The proposed framework processes the received waveform and decouples speech and noise signals, which are fed into two short-time Fourier transform (STFT) convolution 1-D layers that map the waveforms to spectrograms in the complex domain. These speech and noise spectrograms are then used to compute the speech mask loss. The proposed method is evaluated using the TIMIT data set for seen and unseen signal-to-noise ratio conditions. It is shown that the proposed method outperforms the speech enhancement methods that use Deep Neural Network (DNN) based speech enhancement or a Speech Enhancement Generative Adversarial Network (SEGAN).

Index Terms: speech enhancement, generative adversarial network, automatic speech recognition

1. Introduction

Speech enhancement is widely used in communication systems, and it plays a key role in Automatic Speech Recognition (ASR) systems. A wide variety of speech enhancement methods have been developed and refined during the last several decades to improve the quality and intelligibility of the degraded speech signal, including spectral-subtraction algorithms, statistical model-based methods that use maximum-likelihood (ML) estimators, Bayesian estimators, minimum mean squared error (MMSE) methods, subspace algorithms based on single value decomposition and noise-estimation algorithms [1].

More recently, deep-learning techniques have been applied to speech enhancement. In [2], a deep auto-encoder was used for denoising, as well as greedy layered pre-training with a fine-tuning strategy. In [3], a deep neural network (DNN) is used. In the training phase, a DNN-based regression model was trained using the log-power spectral features from pairs of noisy and clean speech data. In [4], a smoothed ideal ratio mask (IRM) was estimated in the Mel frequency domain using deep neural networks. In [5], several training targets were investigated for speech separation. Recent research results show that one can effectively suppress noise using generative adversarial networks (GAN) [6]. In [7], an auto-encoder is leveraged as a generator within the GAN framework to process the received raw speech waveform, which is trained in an end-to-end fashion. In [8], a relativistic GAN is proposed that optimizes the relativistic cost function at its discriminator with a gradient penalty to improve time-domain speech enhancement. In [9], a cycle-consistent speech enhancement (CSE) was introduced that uses an additional inverse mapping network to reconstruct the noisy features from the enhanced ones. In [10, 11], GAN-based algorithms were proposed to operate on spectral features instead of time-domain waveforms. In [12], a time-frequency (T-F) masking-based enhancement framework is introduced, which learns the mask implicitly using a GAN while predicting the clean T-F representation. MetricGAN, which was introduced in [13], optimizes the generator with respect to one or multiple speech enhancement evaluation metrics.

The above-mentioned methods can generally be classified as feature-mapping and mask-learning methods, which are two commonly used deep-learning approaches for single-channel speech enhancement methods for stereo data. Feature mapping approaches [2, 3, 8, 11, 14] enhance the noisy features using a mapping network that minimizes the mean square errors between the enhanced and clean features. Mask-learning approaches [4, 5, 12] estimate the ideal ratio mask or the ideal binary mask, and then use this mask to filter noisy speech signals and reconstruct the clean speech signals. Even though the scale of the masked signal is in the same range as the target signal, one typically sees faster convergence because of the constrained dynamic range. Mask-learning methods usually outperform feature mapping approaches with respect to speech quality [15, 16].

This paper introduces a novel speech enhancement approach that integrates the mask-learning and time-domain feature-mapping methods into one unified framework to take advantage of both approaches. The proposed framework uses a forked GAN structure [11] to extract both speech and noise signals. The generated speech and noise signals are fed into two separate short-time Fourier transform (STFT) convolution 1-D layers to generate the speech and noise spectrograms, which are used to calculate the speech mask. The feature mapping component can preserve the phase information, which is useful to improve the speech quality [17]. The proposed speech enhancement system is shown to perform better than DNN-based and GAN-based speech enhancement systems.

This work is supported by the US Army Medical Research and Materiel Command under Contract No. W81XWH-17-C-0238.

http://dx.doi.org/10.21437/Interspeech.2020-1946

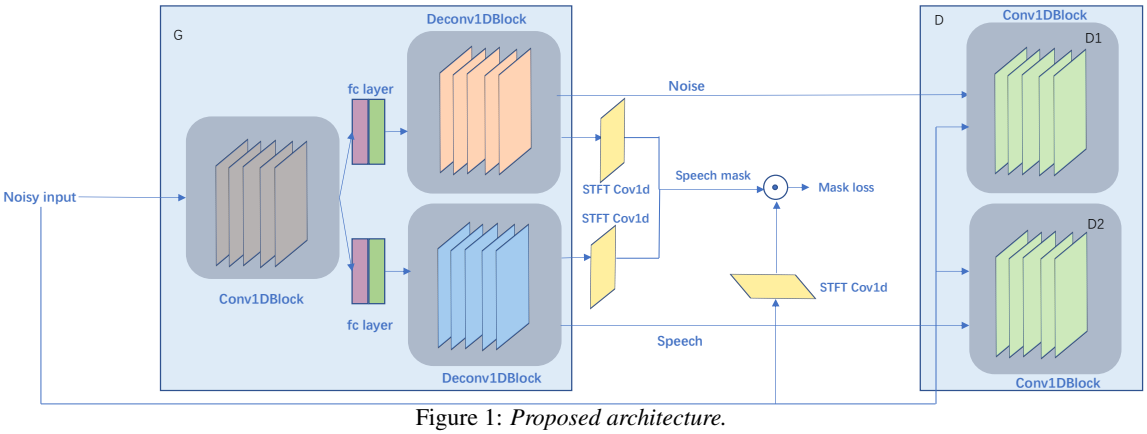


Figure 1: *Proposed architecture.*

2. GAN-based Speech Enhancement

A generalized adversarial network (GAN) consists of a generator network ($\mathbf{G}$) and a discriminator network ($\mathbf{D}$), which uses a alternative mini-max training scheme. Recent studies have shown the potential of exploiting a GAN for speech technology-related applications that aim to learn a suitable mapping function and to accurately reconstruct the enhanced speech while preserving the speech quality and intelligibility.

Consider a noisy speech signal $\boldsymbol{x} = \hat{\boldsymbol{x}} + \boldsymbol{v}$, where $\hat{\boldsymbol{x}}$ is the clean signal and $\boldsymbol{v}$ is the noise signal. Speech enhancement methods aim to reconstruct $\hat{\boldsymbol{x}}$ from $\boldsymbol{x}$. Methods that use GAN-based speech enhancement usually train generator $\mathbf{G}$ to map a noisy speech signal $\boldsymbol{x}$ to its corresponding clean speech signal $\hat{\boldsymbol{x}}$ by minimizing the loss function $\mathcal{L}_{\mathbf{G}}$, which is given by [7]

$$\mathcal{L}_{\mathbf{G}} = \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}),\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[(\mathbf{D}(\mathbf{G}(\boldsymbol{z},\boldsymbol{x}),\boldsymbol{x})-1)^2] + \lambda\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}),\hat{\boldsymbol{x}},\boldsymbol{x}\sim p_{\text{data}}}\|\mathbf{G}(\boldsymbol{z},\boldsymbol{x})-\hat{\boldsymbol{x}}\|_1. \quad (1)$$

The discriminator network $\mathbf{D}$ seeks to distinguish real data from generated data (1 for real, 0 for fake) by minimizing the loss function $\mathcal{L}_{\mathbf{D}}$, which is given by

$$\mathcal{L}_{\mathbf{D}} = \frac{1}{2}\mathbb{E}_{\boldsymbol{x},\tilde{\boldsymbol{x}}\sim p_{\text{data}}(\boldsymbol{x},\hat{\boldsymbol{x}})}[(\mathbf{D}(\hat{\boldsymbol{x}},\boldsymbol{x})-1)^2] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}),\boldsymbol{x}\sim p_{\text{data}}}[\mathbf{D}(\mathbf{G}(\boldsymbol{z},\boldsymbol{x}),\boldsymbol{x})^2]. \quad (2)$$

3. Time-Domain GAN with Mask-Learning

The proposed time-domain GAN with mask-learning method uses a convolution encoder and two parallel de-convolution decoders for speech and noise extraction as illustrated in Fig. 1. It takes a raw waveform as input, and the output of the encoder $\boldsymbol{\Phi}(\boldsymbol{x})$ is fed into two separate fully-connected layers that generate the speech latent representation $\boldsymbol{c}_1$ and the noise latent representation $\boldsymbol{c}_2$, respectively. Each decoder input concatenates an encoder-latent representation with a random vector $\boldsymbol{z}$ that is sampled from a normal distribution $\mathcal{N}(0, I)$, and outputs the predicted time-domain speech signals $\tilde{\boldsymbol{x}} = \boldsymbol{\Psi}_x([\boldsymbol{c}_1, \boldsymbol{z}_1])$ and noise signals $\tilde{\boldsymbol{v}} = \boldsymbol{\Psi}_v([\boldsymbol{c}_2, \boldsymbol{z}_2])$, where $\boldsymbol{\Psi}(\cdot)$ denotes the decoding operation. The generator network also includes skip connections among encoder layers and its homologous decoding layer to avoid losing many low-level details.

Two STFT convolution 1-D layers are used to map the generated speech and noise waveforms to complex spectrograms that include both magnitude and phase components. The magnitude component will be used only. Given a window function ω of length N, the speech complex spectrogram $\tilde{\boldsymbol{X}}_{t,f}$ and the noise complex spectrogram $\tilde{\boldsymbol{V}}_{t,f}$ obtained by STFT can be written as

$$\tilde{\boldsymbol{x}} \xrightarrow{\text{STFT}} \tilde{\boldsymbol{X}}_{t,f} = \sum_{n=0}^{N-1} \tilde{\boldsymbol{x}}\omega\,[n-t]\exp\left(-i\frac{2\pi n}{N}f\right) \quad (3)$$

$$\tilde{\boldsymbol{v}} \xrightarrow{\text{STFT}} \tilde{\boldsymbol{V}}_{t,f} = \sum_{n=0}^{N-1} \tilde{\boldsymbol{v}}\omega\,[n-t]\exp\left(-i\frac{2\pi n}{N}f\right). \quad (4)$$

After having obtained the T-F representation of the enhanced speech and noise, the ideal ratio mask (IRM) and a modified signal approximation (SA) are calculated using

$$\text{IRM} = \sqrt{\frac{\tilde{\boldsymbol{X}}(t,f)^2}{\tilde{\boldsymbol{X}}(t,f)^2 + \tilde{\boldsymbol{V}}(t,f)^2}}, \quad (5)$$

where $\tilde{\boldsymbol{X}}(t,f)^2$ and $\tilde{\boldsymbol{V}}(t,f)^2$ represent the generated speech energy and noise energy with a T-F unit, respectively. Then a signal approximation method is used to train a ratio mask estimator that minimizes the difference between the spectral magnitude of the clean speech and the estimated speech. The mask loss $\mathcal{L}_{\text{mask}}$ is defined as

$$\mathcal{L}_{\text{mask}} = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\|\text{IRM}\odot\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2, \quad (6)$$

where $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are noisy speech and clean speech magnitudes, respectively.

During the training phase, the goal is to minimize the difference between the estimated signal pair $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ and the ground truth signal pair $(\hat{\boldsymbol{x}}, \boldsymbol{v})$ by optimizing the encoder and decoder functions. Similar to SEGAN, the training procedure combines adversarial learning regularized with regression loss. We feed the noisy speech $\boldsymbol{x}$, the clean speech $\hat{\boldsymbol{x}}$, and the additive noise signal $\boldsymbol{v}$ into the proposed framework. In adversarial learning, $\hat{\boldsymbol{x}}$ and $\boldsymbol{v}$ are also used as ground truth for regression in the generator. As such, the generator loss $\mathcal{L}_{\mathbf{G}}$ is the weighted sum of the mask loss, L_1 regular loss and original adversarial loss, which can be written as

$$\mathcal{L}_G = \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}),\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[(\mathbf{D}(\mathbf{G}(\boldsymbol{z},\boldsymbol{x}),\boldsymbol{x})-1)^2\right] + \lambda\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}),\hat{\boldsymbol{x}},\boldsymbol{x}\sim p_{data}}\|\mathbf{G}(\boldsymbol{z},\boldsymbol{x})-\hat{\boldsymbol{x}}\|_1 + \alpha\cdot\mathcal{L}_{\text{mask}} \quad (7)$$

where α denotes the coefficient that controls the contribution of the mask loss function. When $\alpha = 0$, the proposed model is similar to ForkGAN [11], but in the time domain and without noise reduction loss and margin loss. When α is large, the proposed model becomes similar to mask-learning.

Two separate discriminators are adopted in the proposed framework to distinguish between real and fake speech and noise, respectively. During the training process, we sample two pairs of the speech signal: *1)* the real pair of samples, which consists of a clean signal $\hat{\boldsymbol{x}}$ and additive noise $\hat{\boldsymbol{v}}$; *2)* the fake pair of samples, which consists of the enhanced clean speech $\tilde{\boldsymbol{x}}$ and the predicted noise signal $\tilde{\boldsymbol{v}}$. Both signals are conditioned on the noisy speech $\boldsymbol{x}$. Two separate discriminator loss terms are then computed using (2) to update the parameters of the generator.

4. Experiments

Data Sets. The data sets used for the experiments are the TIMIT corpus [18] and the NOISE-100 corpus [19]. The TIMIT corpus is used for clean speech references, and it includes eight major dialects of American-English recorded from 630 speakers, each reading ten phonetically-rich sentences, and partitioned into test and training subsets. The NOISE-100 corpus includes 100 different noise sounds, e.g., animal sounds, and the sound of water. For the training set, a randomly selected noise sound from the NOISE-100 corpus was attached to every silence-added segment with signal-to-noise ratios (SNRs): -3, 0, 3, 6, 9, 12 and 15 dB. In total, this yields 32,340 training samples. We selected 50 sentences from the TIMIT core test and mixed the noise from the NOISE-100 corpus with five SNR conditions (250 sentences in total). For the unseen scenario, we use five unseen SNR conditions at -5, -2, 1, 4, and 7 dB. Note that both seen and unseen conditions were mixed with the same noise from the NOISE-100 corpus.

Baseline Setup. The proposed method is compared with the DNN-based speech enhancement [3], SEGAN [7], and SEGAN+ [1].

DNN-based speech enhancement. Log-spectral features were applied for DNN-based speech enhancement spliced in time by taking a context size of seven frames. In the training stage, a regression DNN model using the mean absolute error (MAE) loss function is trained. The full network topology consists of three hidden layers and 2048 hidden units. The network was trained for 100 epochs using the Adam optimizer with a mini-batch size of 500 and a 20% drop-out in the hidden layers.

SEGAN and SEGAN+. The default parameter settings of the original SEGAN experiments are used, except for the batch size, which is set to 32. Both SEGAN and SEGAN+ take a raw 16,384-sample waveform as input. In SEGAN, $\mathbf{G}$ is composed of 22 1-D strided convolutional layers with filter-width 31 and stride 2. For SEGAN+, this is replaced by a generator comprising 10 1-D convolutional layers and stride 4. The virtual batch-norm (VBN) [20] that is used in SEGAN is replaced by a normal batch normalization in the discriminator.

Setup for the Proposed Method. The proposed model is trained for 100 epochs using an Adam optimizer [21] and a batch size of 32. The proposed approach operates directly on raw audio, which uses a 1-second sliding window with a 50-percent overlap to extract chunks of noisy speech waveforms of 16,384 samples each. A high-frequency pre-emphasis filter with a filter coefficient 0.95 is applied to all input samples during the training and test stages. The generator comprises one encoder and two decoders. Both the encoder and the two decoders consist of five 1-D convolutional layers as shown in Table 1. The speech decoder and the noise decoder have the same structure. Note that the decoder has the skip connections from the encoder part. Two separate fully-connected layers are used for generating speech and noise latent representations. For a short time Fourier transform (STFT) setting, we use a 20 ms Hann window, a 20 ms filter length and a 10 ms hop size. Thus, the input size of STFT is 16,384 and the output is 161×103. For the weight of $\mathcal{L}_{\text{mask}}$, we consider three settings, $\alpha \in \{0, 30, 50\}$, where $\alpha = 0$ means no $\mathcal{L}_{\text{mask}}$ for the training. Two discriminators are used to distinguish fake and real speech and noise, respectively. They both have the same model architecture, which is similar with the encoder in $\mathbf{G}$. We use instance normalization (IN) [22] instead of batch normalization (BN) [23] in the discriminator as we found that IN is slightly better than BN. After convolutional layers, there are three fully connected layers (hidden layer size 256,128,1) with PReLU [24] for binary classification.

Table 1: *Model Structures of the Proposed Method*

layer type		output size
input layer		1×16384
Encoder	conv-1-D	64×4096
	conv-1-D	128×1024
	conv-1-D	256×256
	conv-1-D	512×64
	conv-1-D	1024×16
Fully connected layer		8192
Fully connected layer		16384
Decoders	deconv-1-D	2048×16
	deconv-1-D	1024×64
	deconv-1-D	512×256
	deconv-1-D	256×1024
	deconv-1-D	128×4096
	deconv-1-D	1×16384
STFT conv-1-D		161×103

5. Experimental Results

Evaluation Metrics. Speech enhancement is commonly measured in terms of the perceptual evaluation of speech quality (PESQ) score, see [25, 26], and the Short-Time Objective Intelligibility (STOI) score, see [27]. The PESQ score has a high correlation with subjective evaluation scores, and is mostly used as a compressive objective measure. The PESQ score is computed by comparing the enhanced speech with the clean reference speech, and it ranges from -0.5 to 4.5. The STOI score is highly relevant to human speech intelligibility and the score ranges from 0 to 1.

Performance Results. Measurements were performed using the TIMIT corpus and NOISE-100 corpus to compare the proposed methods with $\alpha \in \{0, 30, 50\}$ and the baseline methods, i.e., the DNN-based method, SEGAN, and SEGAN+. The experimental results are detailed in Table 2. It is shown that the proposed method, with $\alpha = 30$, consistently outperforms the baseline methods for both seen and unseen conditions. The best

[1] https://github.com/santi-pdp/segan_pytorch

Table 2: *Performance of three baseline models and the proposed model. The best values in each column are printed in boldface.*

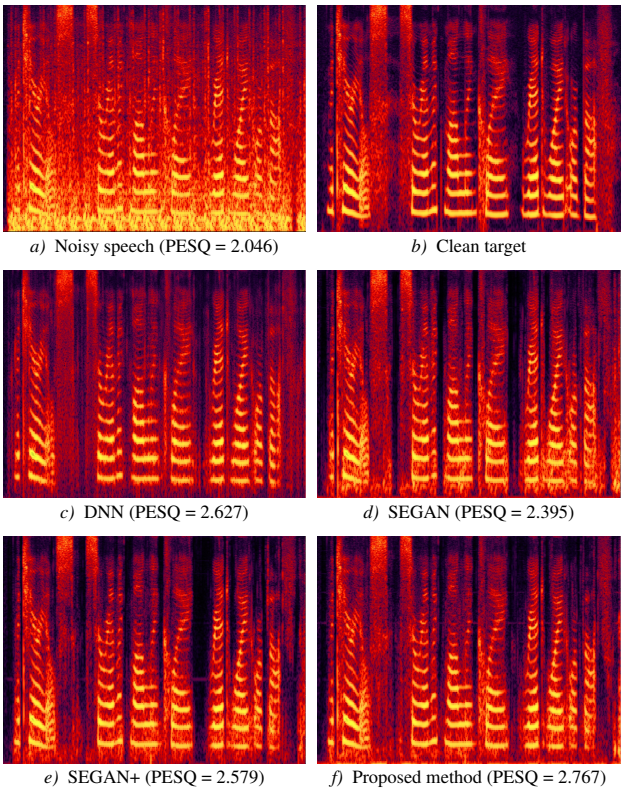
Metrics			w/o SE	DNN	SEGAN	SEGAN+	Proposed method		
							$\alpha = 0$	$\alpha = 30$	$\alpha = 50$
PESQ	seen	-3 dB	1.50 ± 0.33	2.27 ± 0.47	2.15 ± 0.48	2.52 ± 0.38	2.47 ± 0.36	$\mathbf{2.67 \pm 0.33}$	2.55 ± 0.33
		0 dB	1.69 ± 0.32	2.47 ± 0.41	2.37 ± 0.43	2.70 ± 0.38	2.66 ± 0.35	$\mathbf{2.83 \pm 0.34}$	2.73 ± 0.34
		3 dB	1.89 ± 0.31	2.64 ± 0.36	2.56 ± 0.40	2.86 ± 0.38	2.80 ± 0.37	$\mathbf{2.97 \pm 0.36}$	2.89 ± 0.35
		6 dB	2.11 ± 0.30	2.80 ± 0.31	2.75 ± 0.37	3.00 ± 0.38	2.94 ± 0.37	$\mathbf{3.09 \pm 0.40}$	3.01 ± 0.37
		9 dB	2.32 ± 0.29	2.94 ± 0.28	2.93 ± 0.33	3.12 ± 0.36	3.06 ± 0.38	$\mathbf{3.17 \pm 0.46}$	3.11 ± 0.38
		average	1.902	2.624	2.552	2.840	2.786	**2.946**	2.858
	unseen	-5 dB	1.40 ± 0.53	2.25 ± 0.58	2.05 ± 0.50	2.44 ± 0.42	2.34 ± 0.41	$\mathbf{2.54 \pm 0.42}$	2.45 ± 0.39
		-2 dB	1.62 ± 0.48	2.45 ± 0.50	2.26 ± 0.46	2.63 ± 0.40	2.55 ± 0.38	$\mathbf{2.71 \pm 0.41}$	2.64 ± 0.38
		1 dB	1.82 ± 0.48	2.63 ± 0.42	2.46 ± 0.42	2.79 ± 0.40	2.73 ± 0.37	$\mathbf{2.88 \pm 0.39}$	2.81 ± 0.37
		4 dB	2.02 ± 0.47	2.78 ± 0.36	2.64 ± 0.39	2.94 ± 0.38	2.87 ± 0.38	$\mathbf{3.03 \pm 0.37}$	2.95 ± 0.38
		7 dB	2.22 ± 0.46	2.91 ± 0.32	2.82 ± 0.35	3.06 ± 0.36	3.00 ± 0.38	$\mathbf{3.12 \pm 0.38}$	3.06 ± 0.38
		average	1.816	2.604	2.446	2.772	2.698	**2.856**	2.782
STOI	seen	-3 dB	0.7086	0.7576	0.8179	0.8791	0.8634	**0.9056**	0.8886
		0 dB	0.7608	0.7883	0.8577	0.9062	0.8956	**0.9287**	0.9145
		3 dB	0.8094	0.8132	0.8896	0.9261	0.9181	**0.9447**	0.9329
		6 dB	0.8535	0.8331	0.9143	0.9407	0.9351	**0.9549**	0.9464
		9 dB	0.8918	0.8482	0.9336	0.9519	0.947	**0.9577**	0.9559
		average	0.8048	0.8081	0.8826	0.9208	0.9118	**0.9383**	0.9277
	unseen	-5 dB	0.6653	0.7504	0.8037	0.8646	0.8492	**0.8971**	0.8757
		-2 dB	0.7213	0.7845	0.8468	0.8967	0.8872	**0.9235**	0.9067
		1 dB	0.7751	0.8103	0.8809	0.9198	0.9146	**0.9418**	0.9285
		4 dB	0.8245	0.8301	0.9082	0.9367	0.9330	**0.9534**	0.9440
		7 dB	0.8677	0.8450	0.9290	0.9484	0.9463	**0.9605**	0.9546
		average	0.7708	0.8041	0.8737	0.9132	0.9061	**0.9353**	0.9219

baseline method is SEGAN+, and the DNN-based method outperforms SEGAN when using the PESQ metric, and SEGAN performs better than the DNN-based method when using the STOI metric.

Table 2 clearly shows the improvements obtained when applying the mask $\mathcal{L}_{\text{mask}}$ with $\alpha = 30$ relative to the situation where the mask is not used, i.e., for $\alpha = 0$. This shows that the additional mask-based loss helps purify speech signal prediction. For instance, the average PESQ and STOI scores were improved from 2.698 to 2.856 and from 0.9061 to 0.9353 on unseen SNR conditions, respectively. Furthermore, the proposed method outperforms all the baseline systems on both seen and unseen SNR conditions. When compared with SEGAN+, the proposed approach improves the average PESQ from 2.772 to 2.856, and the average STOI from 0.9132 to 0.9353 on unseen SNR conditions. We also notice that if we use a large value for α, corresponding to a strong contribution of $\mathcal{L}_{\text{mask}}$, the performance degrades slightly, because mask-based learning introduces inaccurate information during training due to inaccuracies in the mask estimator. Thus, the loss of mask-based learning and time-domain feature mapping need to be calibrated.

6. Conclusions

In this paper, we propose a novel GAN-based speech enhancement method that integrates mask-learning and feature-mapping. Experimental results with the TIMIT data set show that the proposed approach achieves a better performance for seen and unseen conditions with varying SNR when compared with the baseline systems, i.e., DNN-based speech enhancement, SEGAN and its improved version SEGAN+. We also verified the effectiveness of the proposed mask-based loss. We are currently investigating the use of different mask-based training targets within this framework.



a) Noisy speech (PESQ = 2.046) *b)* Clean target

c) DNN (PESQ = 2.627) *d)* SEGAN (PESQ = 2.395)

e) SEGAN+ (PESQ = 2.579) *f)* Proposed method (PESQ = 2.767)

Figure 2: *Spectrograms of a sample input mixed with N21 noise, where the SNR is equal to 1 dB.*

7. References

[1] P. C. Loizou, *Speech Enhancement*, 2nd ed. Boca Raton, FL: CRC Press, 2013.

[2] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 436–440.

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

[4] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Proc.*, Vancouver, BC, May 2013, pp. 7092–7096.

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

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial sets," in *Proc. Conf. on Neural Inform. Proc. Sys.*, Montréal, QC, Dec. 2014.

[7] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," *ArXiV:1703.09452*, 2017.

[8] D. Baby and S. Verhulst, "SERGAN: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Proc."*, Brighton, United Kingdom, May 2019, pp. 106–110.

[9] Z. Meng, J. Li, Y. Gong *et al.*, "Cycle-consistent speech enhancement," *arXiv preprint arXiv:1809.02253*, 2018.

[10] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Proc.*, Calgary, AB, Apr. 2018, pp. 5024–5028.

[11] J. Lin, S. Niu, Z. Wei, X. Lan, A. J. van Wijngaarden, M. C. Smith, and K.-C. Wang, "Speech enhancement using forked generative adversarial networks with spectral subtraction," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 3163–3167.

[12] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Proc.*, Calgary, AB, Apr. 2018, pp. 5039–5043.

[13] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," ArXiV, Pre-Print 1905.04874, May 2019.

[14] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Proc. Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[15] Z. Chen, Y. Huang, J. Li, and Y. Gong, "Improving mask learning based speech enhancement system with restoration layers and residual connection," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 3632–3636.

[16] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 826–835, 2014.

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

[18] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," Linguistic Data Consortium, Tech. Rep., 1993.

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

[20] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

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

[22] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int'l Conf. Computer Vision*, Las Condes, Chile, Dec. 2015, pp. 1026–1034.

[25] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Proc.*, Salt Lake City, UT, May 2001, pp. 749–752.

[26] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T recommendation P.862, Feb. 2001.

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