TIME-FREQUENCY MASKING-BASED SPEECH ENHANCEMENT USING GENERATIVE ADVERSARIAL NETWORK

Meet H. Soni, Neil Shah, and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India

{meet_soni, neil_shah, and hemant_patil}@daiict.ac.in

ABSTRACT

The success of time-frequency (T-F) mask-based approaches is dependent on the accuracy of predicted mask given the noisy spectral features. The state-of-the-art methods in T-F masking-based enhancement employ Deep Neural Network (DNN) to predict mask. Recently, Generative Adversarial Networks (GAN) are gaining popularity instead of maximum likelihood (ML)-based optimization of deep learning architectures. In this paper, we propose to exploit GAN in T-F masking-based enhancement framework. We present the viable strategy to use GAN in such application by modifying the existing approach. To achieve this, we use a method that learns the mask implicitly while predicting the clean T-F representation. Moreover, we show the failure of vanilla GAN in predicting the accurate mask and propose a regularized objective function with the use of Mean Square Error (MSE) between predicted and target spectrum to overcome it. The objective evaluation of the proposed method shows the improvement in the accurate mask prediction, as against the state-of-the-art ML-based optimization techniques. The proposed system significantly improves over a recent GAN-based speech enhancement system in improving speech quality, while maintaining a better trade-off between less speech distortion and more effective removal of background interferences present in the noisy mixture.

Index Terms— Task-dependent masking, speech enhancement, generative adversarial networks

1. INTRODUCTION

The objective of speech enhancement is to improve the speech intelligibility and quality given the noisy mixture [1]. Speech enhancement is essential in many applications to generate the robust speech-specific features and/or the enhanced speech waveform. Such robust features find their applications in speech recognition and speaker identification task [2]. The enhanced speech can be used in hearing aid devices and cochlear implant (CI) designs [3].

This work was supported in part by MeitY, Govt. of India, through a consortium projects ASR Phase-II, and in part by the authorities of DA-IICT, Gandhinagar, India.

Time-Frequency (T-F) masking-based approaches employing supervised learning are state-of-the-art techniques in the enhancement and source separation problems [4–8]. The aim of supervised learning is to predict the accurate T-F mask given the noisy mixture. In such approaches, a Deep Neural Network (DNN) is often used to predict the T-F mask using the features extracted from the noisy mixture [5–10]. Currently, all such approaches use Maximum Likelihood (ML)-based optimization criteria to predict the T-F mask or clean T-F representation while predicting the mask implicitly. ML-based optimization criteria puts prior assumptions on data distribution (such as, Minimum Mean Square Error (MMSE) objective function assumes the output variables to be Gaussian) which may not be valid for the given data. Often such assumptions prevent the network to learn perceptually optimal network parameters for various several speech technology applications. For T-F masking-based approaches, the difference between the performance of the oracle mask and the predicted mask indicates the need of better objective function to perceptually optimize the network parameters [2]. Generative Adversarial Networks (GAN) provides one such alternative of ML-based optimization criteria [11]. In this paper, we propose to exploit GAN for T-F masking-based speech enhancement task. We have shown that the objective function of vanilla GAN (v-GAN) is not sufficient to predict the T-F mask accurately and we have modified the objective function to address this limitation. Moreover, the proposed GAN-based speech enhancement framework is generalizable to any T-F representation.

1.1. Recent Work

GAN is a deep learning (DL) architecture [11], which is a well-established generative modeling technique in the field of computer vision [12–14]. Recently, the use of GAN is gaining popularity in speech technology applications that require accurate reconstruction of speech. The use of GAN has shown improvements over ML-based techniques in voice conversion (VC) [15, 16] and speech enhancement (SE) task [17, 18]. The conditional GAN (cGANs) architecture proposed in [17], uses a Pixel-to-Pixel (Pix2Pix) framework for SE task, by learning the mapping function between the noisy

speech and clean speech spectrogram. A GAN-based post-filter for Short-Time Fourier Transform (STFT) spectrograms proposed in [19] reconstructs the spectrogram that preserves the finer structures and resembles the true data, even in the high-dimensional STFT-domain. In addition, a speech enhancement GAN (SEGAN) [18] have shown a promising result for end-to-end speech enhancement task in an adversarial framework. These approaches directly predicts the spectrum [17] or raw samples [18] of the clean speech. Such approach may not be suitable for applications, such as source separation, where T-F masking-based methods are proven to be a better approach. In this paper, we present a viable framework to exploit GAN in T-F masking-based enhancement.

2. PROPOSED FRAMEWORK FOR T-F MASKING-BASED ENHANCEMENT

2.1. Generative Adversarial Networks (GANs)

The aim of the generative model is to produce the samples that resemble the samples generated from the data distribution $\mathcal{X}$. GAN is a generative model that learns the mapping between the samples y from some prior distribution $\mathcal{Y}$ to samples x belonging to $\mathcal{X}$. The G network is responsible for learning the mapping function in an adversarial framework along with a D network. Typically, a D network is a binary classifier with input as real samples coming from $\mathcal{X}$ and the fake samples generated by G. The adversarial characteristics of the GAN forces the D network to maximize the likelihood of the samples coming from $\mathcal{X}$ as real, whereas minimizing the likelihood of the samples coming from the model distribution $\hat{\mathcal{X}}$ (output of G) as fake. As training proceeds, the G network adjust its parameters by generating realistic samples at its output, as a result the generated samples closely follows $\mathcal{X}$, leaving the D network unable to differentiate between the true and fake distributions. This objective function can be formulated as [11]:

$$\min_D V(D) = -\mathbb{E}_{x\sim\mathcal{X}}[\log D(x)] - \mathbb{E}_{y\sim\mathcal{Y}}[1-\log(D(G(y)))], \tag{1}$$

$$\min_G V(G) = -\mathbb{E}_{y\sim\mathcal{Y}}[\log D(G(y))], \tag{2}$$

where $\mathbb{E}_{x\sim\mathcal{X}}$ denotes the expectation over all the samples x coming from the distribution $\mathcal{X}$.

3. T-F MASKING USING GAN

GAN can be used in T-F masking-based approaches, where the objective of the G network is to generate T-F mask or clean T-F representation and the objective of the D network is to differentiate between the actual T-F mask or clean T-F representation and the one generated by G. We employ a method in which G is trained to directly predict the clean representation while learning the mask-like representation implicitly [9]. Such method referred to as task-dependent masking, has shown to give better performance than the directly predicting T-F mask using the DNN [9]. Moreover, the recent studies in GAN have shown that the D network can accurately learn to discriminate between the clean and noisy spectrum [19]. This method generalizes well to the different feature space, such as filterbank energies used in speech recognition, Short-Time Fourier Transform (STFT) spectrum, and Gammatone spectrum.

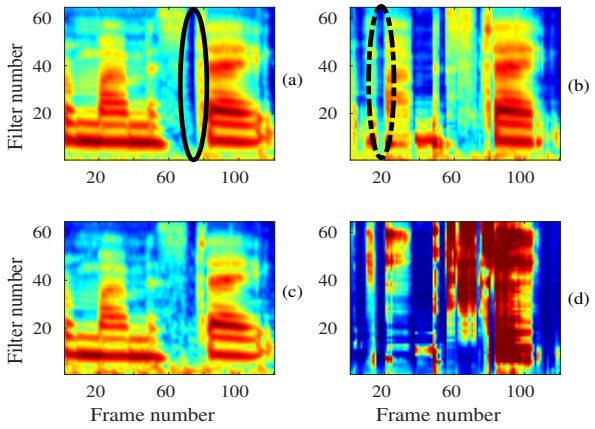

Fig. 1. GAN fails to properly predict the mask (a) clean T-F representation: the solid-circle region shows the silence frame, (b) enhanced T-F representation: the dotted-circle shows the predicted frame where GAN fails, (c) noisy T-F representation, and (d) predicted mask.

Inspired by [9], we propose to optimize error between the log T-F representation of clean and enhanced speech, while learning the mask-like representation at the output layer of the network. If the output of the last layer of the network is m, then the objective function for the network parameter optimization can be written as:

$$\begin{aligned} x &= \log(c), \\ t &= \log(h \circ m), \\ J &= \frac{1}{2}||t-x||^2, \end{aligned} \tag{3}$$

where c is the clean T-F representation, h is noisy T-F representation, $\circ$ denotes the elementwise multiplication, and J is the objective function to train the network parameters. Using this objective function, the gradient equations for backpropagation algorithm can be easily written. Hence, the network will learn a T-F mask which, if multiplied with the noisy T-F representation, will produce clean T-F representation. If m is constrained to have values between 0 to 1, the T-F mask learned by the network should resemble (but not exactly) to

the Ideal Ratio Mask (IRM) [9].

The GAN can be easily employed in such framework. Here, the objective of the G network is to learn the accurate clean T-F representation while learning the mask implicitly. The D network is trained to differentiate between the clean T-F representation and output generated by G. The initial experiments using vanilla GAN suggests that this framework, while viable intuitively, fails to learn the T-F mask accurately. Fig. 1 shows one instance of such failure. The dotted circle in Fig. 1 (b) shows the area where GAN is not able to predict the mask accurately. However, the enhanced T-F representation (Gammatone spectrum) of the region resembles the region of the clean T-F representation in Fig. 1 (a) showed by the solid circle. The output of G is not accurate for the given frame, while it still belongs to the distribution of clean T-F representation ($\mathcal{X}$). Hence, the D is not able to differentiate it as fake representation and learning fails. The cost of D is also observed to be low at such instances.

One possible solution to prevent this is to *regularize* the objective function. The G network is able to fool D by generating enhanced T-F representation belonging to some other frame, which still resembles the clean T-F distribution. Hence, we use the MMSE error between the predicted and the clean T-F representation in addition to vanilla GAN objective function of G. The modified objective function can be written as:

$$\min_G V(G) = -\mathbb{E}_{y\sim\mathcal{Y}}[\log(D(G(y)))] + \frac{1}{2}\mathbb{E}_{x\sim\mathcal{X},y\sim\mathcal{Y}}[\log(x) - \log(G(y))]^2. \quad (4)$$

The similar objective function is used in [17] and [18]. They have used the L_1 norm between network output and the target to regularize the training. However, they have not justified the need of such regularization in the context of speech applications. We use the MMSE error since we are comparing our results with the network trained with the MMSE criteria.

4. EXPERIMENTAL SETUP

4.1. Database

The proposed algorithm was evaluated on the dataset released by Valentini et. al. [20]. The training and testing set have mismatched condition. The dataset comprises of 30 speakers from the Voice Bank corpus [21]. The training set contains 28 native English speakers with around 400 sentences for the clean and noisy set, sampled at 48 kHz. The test set contains 2 native English speakers with around 400 sentences for the clean and noisy set. The total of 40 different noisy conditions with 10 types of noise (2 artificial and 8 from Demand database [22]) and 4 signal-to-noise ratio (SNR) each (15, 10, 5, and 0 dB) are considered for the noisy training set. There are around 10 different sentences per training speaker in each condition. To make the test set, a total of 20 different noisy conditions with 5 types of noise (all from the Demand database) and 4 SNR each (17.5, 12.5, 7.5, and 2.5 dB) are considered. Training set contains 11,572 utterances, while 824 utterances are available in the test set.

4.2. Network architecture

We trained three networks to compare the results. The first network is a DNN which is trained using the MMSE criteria between enhanced and clean T-F spectrum. The second network is vanilla GAN (v-GAN) and the third network being GAN with MMSE regularization (MMSE-GAN). In v-GAN and MMSE-GAN, the G network was identical to the DNN network. DNN and G of both GAN and MMSE-GAN had three hidden layers. Each layer had 512 units with Rectified Linear Unit (ReLU) activation. The output layer had 64 units to predict T-F mask implicitly. Sigmoid activation was used to limit the output mask values between 0 to 1. The D network of GAN and MMSE-GAN networks also had three hidden layers with 512 units in each layer. However, the units has tanh activation function. The output layer had single unit with sigmoid activation. All the three models were trained for 30 epochs with Adam optimizer [24] and a learning rate of 0.001, using a batch size of 1000.

To prepare the input-output pair to train the network, Gammatone spectral features were extracted from the speech signals. The original utterances of the database were downsampled from 48 kHz to 16 kHz. Pre-emphasis with the factor 0.95 was performed. Then, 64-channel Gammatone spectrum was computed with 20 ms Hamming window and 10 ms overlap between consecutive frames. The input to the network was 7 frames (3 left and 3 right) context of log-Gammatone spectrum. The networks were trained to predict the clean log-Gammatone spectrum as T-F representation, while learning the mask implicitly. Out of total 11572 training utterances, 11000 random utterances were used to train the networks and remaining 572 utterances were taken as validation set. Once the network is trained, the model with the least MSE on validation dataset was chosen and testing was performed.

4.3. Results

Fig. 2 shows the predicted masks for three different architectures. The visual inspection shows that the mask predicted by the MMSE-GAN is significantly better than the other two networks. Especially, the mask predicted by MMSE-GAN preserves the finer structure in the predicted mask. To evaluate the performance over the entire database, the quality of the enhanced speech is computed using various objective measures. CSIG measured from 1 to 5, predicts the mean opinion score (MOS) of the signal distortion, considering only to the speech signal [25]. CBAK and CMOS measured from 1 to 5, predicts the extent of background interferences in the speech and the overall effect, respectively [25]. Perceptual

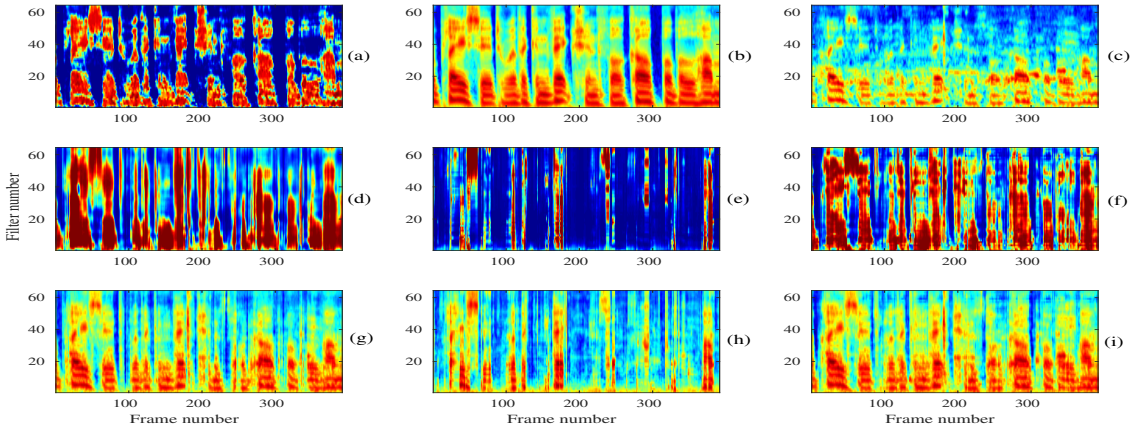

Fig. 2. (a) Oracle mask, Gammatone spectrum of (b) clean speech, (c) noisy speech. Predicted mask using (d) RMSE-DNN, (e) GAN, (f) RMSE-GAN. Gammatone spectrum of reconstructed speech using (g) DNN, (h) GAN, (i) MMSE-GAN.

Table 1. Performance comparisons between the noisy signal, DNN, MMSE-GAN, GAN, SEGAN, and the Wiener filter-based enhancement

Metric	Noisy	DNN	v-GAN	MMSE-GAN	SEGAN [18]	Wiener [23]
CSIG	3.35	3.73	2.48	**3.80**	3.48	3.23
CBAK	2.44	3.09	2.64	**3.12**	2.94	2.68
CMOS	2.63	3.09	1.91	**3.14**	2.8	2.67
PESQ	1.97	2.49	1.41	**2.53**	2.16	2.22
STOI	0.91	**0.93**	0.79	**0.93**	**0.93**	-

Evaluation of Speech Quality (PESQ) measured from (-0.5 to 4.5), stands for perceptual evaluation of speech quality is a wideband version as recommended in ITU-T P.862.2 [26] to assess the voice quality in the speech. All these metrics are computed using the implementation shown in [1]. Moreover, to show the improvement in speech intelligibility, we also calculated Short-Time Objective Intelligibility (STOI) measure [27].

Table 1 shows the metric scores for the different architectures. We compare the results of our approach with the existing speech enhancement algorithms, such as SEGAN [18] and Weiner filter-based method [23]. The results for SEGAN and Wiener filter-based methods are directly taken from the [18], since the same database and evaluation metrics are used in their work. The quality scores suggest that v-GAN is not able to improve speech quality due to inaccurate prediction of the mask. While MMSE-GAN gives significant performance improvement over DNN, especially in improving signal and overall quality. It has to be noted that this improvement is solely due to employing GAN in optimizing the network parameters, since *all* the other conditions were similar while training the DNN and G. Moreover, the comparison with SEGAN architecture suggest that T-F masking-based approaches are better for speech enhancement, at least in terms of metrics for evaluating the objective quality. The STOI scores suggest that the intelligibility of the speech signals using DNN, MMSE-GAN, and SEGAN is almost similar.

5. SUMMARY AND CONCLUSIONS

In this study, we proposed and analyzed a framework for T-F mask estimation using a Generative Adversarial Networks (GAN). In this study, a DNN-based GAN is employed for mask estimation. We show that the vanilla GAN is insufficient to learn the accurate spectral mapping, given the noisy T-F representation. To that effect, we establish the need of MMSE regularization in the GAN framework and have shown its viability. Results show that the proposed framework estimates the mask more accurately than the DNN trained using MMSE criteria. The objective measures also dictate the improvement in the performance by using adversarial training. The presented framework can be improved in many ways. The use of MSE error in the discriminator network instead of cross-entropy is proven to be better. The L_1 norm can be used to regularize the cost function of generator. Convolutional Neural Network (CNN)-based networks can be used in the some manner, since they are proven to perform better in GAN.

6. REFERENCES

[1] P. C. Loizou, "Speech Enhancement: Theory and Practice," 2^{nd} ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[2] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. W. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *12th International Conference on Latent Variable Analysis and Signal Separation LVA/ICA*, Liberec, Czech Republic, 2015, pp. 91–99.

[3] L.-P. Yang and Q.-J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *The Journal of the Acoustical Society of America (JASA)*, vol. 117, no. 3, pp. 1001–1004, 2005.

[4] Y. Wang and D. Wang, "A structure-preserving training target for supervised speech separation," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 6127–6131.

[5] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America (JASA)*, vol. 139, no. 5, pp. 2604–2612, 2016.

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

[7] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, British Columbia, Canada, 2013, pp. 7092–7096.

[8] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation*. Springer, 2014, pp. 349–368.

[9] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 4390–4394.

[10] E. Grais, M. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[11] 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 (NIPS)*, Montreal, Canada, 2014, pp. 2672–2680.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[14] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Nevada, USA, 2016, pp. 2536–2544.

[15] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," *INTERSPEECH*, pp. 1283–1287, 2017.

[16] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3364–3368.

[17] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2008–2012.

[18] S. Pascual, A. Bonafonte, and J. Serr, "Segan: Speech enhancement generative adversarial network," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3642–3646.

[19] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for stft spectrograms," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3389–3393.

[20] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," http://dx.doi.org/10.7488/ds/1356, [Publicly available Online; Last accessed 25-September-2017].

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

[22] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America (JASA)*, vol. 133, no. 5, pp. 3591–3591, 2013.

[23] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 26, no. 3, pp. 197–210, 1978.

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

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

[26] "P.862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *Geneva: International Telecommunication Union*, 2007.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010, pp. 4214–4217.