Conditional Generative Adversarial Networks for Speech Enhancement and Noise-Robust Speaker Verification

Daniel Michelsanti and Zheng-Hua Tan

Department of Electronic Systems, Aalborg University, Denmark

dmiche15@student.aau.dk, zt@es.aau.dk

Abstract

Improving speech system performance in noisy environments remains a challenging task, and speech enhancement (SE) is one of the effective techniques to solve the problem. Motivated by the promising results of generative adversarial networks (GANs) in a variety of image processing tasks, we explore the potential of conditional GANs (cGANs) for SE, and in particular, we make use of the image processing framework proposed by Isola et al. [1] to learn a mapping from the spectrogram of noisy speech to an enhanced counterpart. The SE cGAN consists of two networks, trained in an adversarial manner: a generator that tries to enhance the input noisy spectrogram, and a discriminator that tries to distinguish between enhanced spectrograms provided by the generator and clean ones from the database using the noisy spectrogram as a condition. We evaluate the performance of the cGAN method in terms of perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and equal error rate (EER) of speaker verification (an example application). Experimental results show that the cGAN method overall outperforms the classical short-time spectral amplitude minimum mean square error (STSA-MMSE) SE algorithm, and is comparable to a deep neural network-based SE approach (DNN-SE).

Index Terms: generative adversarial networks, speech enhancement, speaker verification

1. Introduction

Dealing with degraded speech signals is a challenging yet important task in many applications, e.g. automatic speaker verification (ASV) [2], speech recognition [3], mobile communications and hearing assistive devices [4, 5, 6]. When the receiver is a human user, the objective of SE is to improve quality and intelligibility of noisy speech signals. When it is an automatic speech system, the goal is to improve the noise-robustness of the system, e.g. to reduce the EERs of an ASV system under adverse conditions. In the past, this problem has been tackled with statistical methods like Wiener filter and STSA-MMSE [7]. Lately, deep learning methods have been used, such as DNNs [6, 8], deep autoencoders (DAEs) [5], and convolutional neural networks (CNNs) [9]. However, to our knowledge, no one has tried to use GANs for SE yet.

GANs are a framework recently introduced by Goodfellow et al. [10], which consists of a generative model, or generator (G), and a discriminative model, or discriminator (D), that play a min-max game between each other. In particular, G tries to fool D which is trained to distinguish the output of G from the real data. The architectures used in most of the works today [11] are based on deep convolutional GAN (DCGAN) [12] that successfully tackles training instability issues when GANs are applied to high resolution images. Three key ideas are used to accomplish this goal. First, batch normalization [13] is applied to most of the layers. Then, the networks are designed to have no pooling layers as done in [14]. Finally, the training is performed adopting the Adam optimizer [15].

So far GANs have been successfully applied to a variety of computer vision and image processing tasks [1, 12, 16, 17]. However, their adoption for speech-related tasks is rare with one exception in [18], in which the authors of the report applied a deep visual analogy network [19] as a generator of a GAN for voice conversion, and the results are presented as example audio files without speech quality or intelligibility or other measures. In a related field, for music, the GAN concept was applied to train a recurrent neural network for classical music generation [20].

Very recently, a general-purpose cGAN framework called Pix2Pix was proposed for image-to-image translation [1]. Motivated by its successful deployment on several tasks, we adapt the framework in this work, aiming to explore the potential of cGANs for SE, as part of the overall goal of investigating the feasibility and performance of GANs for speech processing. Specifically, we use Pix2Pix to learn a mapping between noisy and clean speech spectrograms as well as to learn a loss function for training the mapping.

2. Pix2Pix framework for speech enhancement

In GANs, G represents a mapping function from a random noise vector $\mathbf{z}$ to an output sample $G(\mathbf{z})$, ideally indistinguishable from the real data $\mathbf{x}$ [10]. In cGANs, both G and D are conditioned on some extra information $\mathbf{y}$ [1], and they are trained following a min-max game with the objective:

$$L(D,G) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\, p_{data}(\mathbf{x},\mathbf{y})}[\log(D(\mathbf{x},\mathbf{y}))] + \mathbb{E}_{\mathbf{z}\sim\, p_{\mathbf{z}}(\mathbf{z}),\mathbf{y}\sim\, p_{data}(\mathbf{y})}[\log(1 - D(G(\mathbf{z},\mathbf{y}),\mathbf{y}))]. \quad (1)$$

Pix2Pix differs from other cGAN works, like [21], because it does not use $\mathbf{z}$. Isola et al. [1] report that adding a Gaussian noise as an input to G, as done in [22], was not effective. Hence, they introduce noise in the form of dropout, but this technique failed to produce stochastic output. However, we are more interested in an accurate mapping between a noisy spectrogram and a clean one than a cGAN able to capture the full entropy of the distribution it models, so this represents a minor issue. Figure 1 shows how the data and the condition are used during training in the particular case of this paper.

In addition to the adversarial loss $L(D,G)$ that is learned from the data, Pix2Pix utilizes also L1 distance between the output of G and the ground truth. The choice of combining different losses, like L2 distance [23] or perceptual losses for a specific task [16, 17], has been shown to be beneficial. In Pix2Pix, L1 distance is preferred to L2 because it encourages less blurring [1] and it tends to generalize better if compared to perceptual losses.

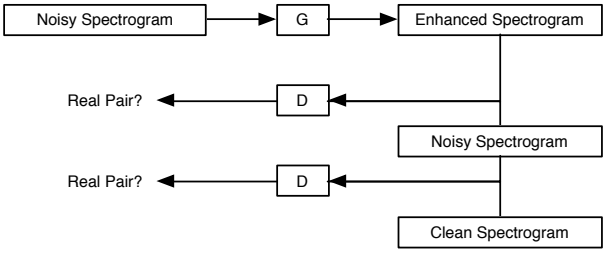

Figure 1: *Generator (G) and discriminator (D) in the Pix2Pix framework for speech enhancement. G generates an enhanced spectrogram from a noisy input by fooling D, which tries to classify a spectrogram as clean or enhanced, conditioned on the respective noisy spectrogram.*

Furthermore, G and D, adapted from [12], are a U-Net [24] and a PatchGAN, respectively. Since in image-to-image translation tasks, the input and the output of G share the same structure, G is an encoder-decoder where each feature map of the decoder layers is concatenated with its mirrored counterpart from the encoder to avoid that the innermost layer represents a bottleneck for the information flow. Besides, D is built to model the high frequencies of the data, as the low frequency structure is captured by the L1 loss. This is achieved by considering local image patches. In particular, D is applied convolutionally across the image to classify each patch as real or fake. Then, the obtained scores are averaged together to get a single output. This architecture has the advantage of being smaller and can be applied on images of different sizes [1]. When the patch size of D has the same size of the input image, D is equivalent to a classical GAN discriminator.

Our Pix2Pix implementation is based on [25], with G that gets a 256×256 1-channel image, while D a 256×256 2-channel image. The main differences with the original framework are the adoption of 5×5 filters in the convolutional layers, and the last layer of D which is flattened and fed into a single sigmoid output as in [12].

2.1. Preprocessing and training

For speech signals with a sample rate of 16 kHz, we computed a time-frequency (T-F) representation using a 512-point short time Fourier transform (STFT) with a hamming window size of 32 ms and a hop size of 16 ms. In this way, the frequency resolution is 16 kHz / 512 = 31.25 Hz per frequency bin. We considered only the 257-point STFT magnitude vectors which cover the positive frequencies due to symmetry. Our generator G accepts $256 \times 256 \times 1$ input, so for training we concatenated all the speech signals and then split the spectrogram every 256 frames, while for testing we zero-padded the spectrogram of each test sample in order to have the number of frames equal to a multiple of 256 and then performed the split accordingly. We also removed the last row of the spectrogram, which is a choice that has a negligible impact since it represents only the highest 31.25 Hz band of the signal, but this allows us to have a power-of-2 input size which makes the design of G and D easier. Before the data are fed to our system, they are also normalized to be in the range $[-1, 1]$.

We trained the GANs using stochastic gradient descent (SGD) and adopting the Adam optimizer, for 10 epochs with a batch size of 1 according to [1], updating G twice per each iteration to avoid a fast convergence of D [25]. The networks' weights have been initialized from a normal distribution with zero mean and a standard deviation of 0.02 [1]. The L1 loss has been added to the GAN loss using a scaling factor of 100 [1].

To enhance a speech signal with Pix2Pix, we first compute the T-F representation of it, and then we forward propagate the spectrogram magnitude through G. Finally, we reconstruct the signal with the inverse STFT using the phase of the noisy input.

3. Experiments

3.1. Evaluation metrics

The performance of our system is evaluated in terms of PESQ [26] (in particular the wide-band extension [27]), STOI [28], and EER of ASV. PESQ and STOI have been chosen as they are the most used estimators of speech quality and speech intelligibility, respectively. The implementations used in this paper are from [7] for PESQ and from [28] for STOI.

Regarding the ASV evaluation, we use the classical Gaussian Mixture Model - Universal Background Model (GMM-UBM) framework [29], which is suitable for short utterances as in this work. We first built a general model, UBM, which is a GMM trained with an expectation-maximization algorithm using a large amount of speech data not belonging to the target speakers. Then, a target speaker model for each specific passphrase and each speaker was derived by maximum a posteriori adaptation of the UBM. The approach of adapting UBM is used in order to have a well-trained model for a speaker even when there is no much data available. At this point, for a test utterance we calculate the log likelihood ratio between the claimant speaker model and the UBM. The features extracted from the speech data are 57-dimensional mel-frequency cepstral coefficients (MFCCs), and the GMM mixture number is 512.

3.2. Baseline methods

We compare the results of our approach with other two methods we consider as baselines: STSA-MMSE and an Ideal Ratio Mask (IRM) based DNN-SE algorithms.

STSA-MMSE is a statistical-based SE technique, where the a priori signal to noise ratio (SNR) is estimated with the Decision-Directed approach [30] and the noise Power Spectral Density (PSD) is estimated with the noise PSD tracker in [31]. The noise PSD estimate is initialized with the first 1000 samples of each utterance, assumed to be a speech-free region.

For the DNN-SE algorithm, we use the same procedure and parameters of [6]. The IRM is estimated by using a DNN with three hidden layers of 1024 units each, and an output layer with 64 units. The input of the DNN is a 1845-dimensional feature vector, which is a robust representation of a frame that combines MFCCs, amplitude modulation spectrogram, relative spectral transform - perceptual linear prediction (RASTA-PLP), and gammatone filter bank energies, with their delta and double delta for a context of 2 past and 2 future frames. The training label is represented by the IRM, which is computed as in [32] from the T-F representation based on a gammatone filter bank with 64 filters linearly spaced on a Mel frequency scale and with a bandwidth equal to one equivalent rectangular bandwidth [33]. The system is trained for 30 epochs with SGD, using the mean square error as error function and a batch size of 1024. In order to enhance a test signal, the DNN provides an estimation of the IRM which is applied to the T-F representation of the noisy signal. Finally, the time domain signal is synthesized.

3.3. Datasets

We use two corpora, TIMIT [34] and RSR2015 [35], as follows:

- Set 1 (TIMIT) - 4380 utterances of male speakers are used for UBM training.
- Set 2 (RSR2015) - Text ID from 2 to 30 of sessions 1, 4, and 7 for 50 male speakers (from m051 to m100) are selected to train Pix2Pix and DNN-SE.
- Set 3 (RSR2015) - Text ID 1 of sessions 1, 4, and 7 for 49 male speakers (from m002 to m050) are used to train the speaker models.
- Set 4 (RSR2015) - Sessions 2, 3, 5, 6, 8, and 9 of the same text ID and speakers used for training the models, are selected for evaluation.

The choice of RSR2015 as the main database for training and testing can be seen as a compromise, because we were interested in the evaluation of an ASV system, which provides another objective measure of the performance, and RSR2015 is widely used for this task.

We used 5 different noise types to simulate real-life conditions: Babble, obtained by adding 6 random speech samples from the Librispeech corpus [36]; white Gaussian noise generated in MATLAB; Cantine, recorded by the authors; Market and Airplane, collected by Fondazione Ugo Bordoni (FUB) and available on request from the OCTAVE project [37]. Noise data, which were added to the utterances in Set 2, 3, and 4 at different SNR values, used for training and testing are different.

3.4. Setup

Inspired by [2], we investigate two different kinds of Pix2Pix-based SE front-ends: 5 noise specific front-ends (NS-Pix2Pix), each of them trained on only one type of noise, and 1 noise general front-end (NG-Pix2Pix), trained on all types of noise. The same has been done for the DNN-SE front-ends, obtaining 5 noise specific front-ends (NS-DNN) and 1 noise general front-end (NG-DNN). For training, we add noise to clean speech at two different SNRs, 10 and 20 dB. With higher SNR it should be easier to train a G able to capture the underlying structure of the noisy input and generate a clean spectrogram, but a test with lower SNRs for training is worth to explore in the future. For testing, results for 5 different SNR conditions are reported: 0, 5, 10, 15, and 20 dB, as is commonly done for ASV, but an interesting future work is to test on lower SNRs, particularly for intelligibility evaluation. In addition, to find the behavior of the front-ends on noise free conditions, ASV performance on enhanced clean speech data is also reported.

In all the tests, the performance of the following front-ends are presented: No enhancement (when no SE algorithm is used on noisy data), STSA-MMSE, NS-DNN, NS-Pix2Pix, NG-DNN, and NG-Pix2Pix. In total, three different tests have been conducted:

- Test 1 - In the first test, we compute PESQ and STOI for the different front-ends to estimate speech quality and intelligibility.
- Test 2 - In the second test, the ASV system is trained with enhanced clean speech (except for the No enhancement front-end where clean speech is used) and tested on the 5 types of noise.
- Test 3 - The last test is performed to evaluate the effects of a multi-condition training on ASV. For No enhancement, STSA-MMSE, NS-DNN, and NS-Pix2Pix the speaker models are built from enhanced clean speech and one kind of enhanced noisy speech, while for NG-DNN and NG-Pix2Pix all kinds of noise are used.

4. Results and Discussion

The results of Test 1 are shown in Table 1. It is observed that the average PESQ scores of NS-Pix2Pix and NG-Pix2Pix are always better than the other front-ends. The best performance improvement is achieved between 5 and 15 dB SNR regardless of the noise type. At 20 dB, our approach outperforms the others on Market and White noises, but for Airplane noise STSA-MMSE is the best one, while for Babble and Cantine noises the absence of enhancement is superior indicating that all the SE techniques introduce an amount of distortion surpassing the benefit of noise reduction. At 0 dB, NG-Pix2Pix generally outperforms the noise specific version with an exception (Market noise) and its scores are close to DNN-SE ones.

In terms of STOI, Pix2Pix front-ends perform similarly to STSA-MMSE. However, DNN-SE front-ends are superior in almost all the conditions, even though Pix2Pix front-ends achieve the same or very close results in some situations, e.g. low SNRs for Cantine and Market noises. At 20 dB, we observe the same behavior as the PESQ scores, where the evaluation of not enhanced signals gives a better outcome.

Table 1: *PESQ and STOI performance for the 5 front-ends: No enhancement (a), STSA-MMSE (b), NS-DNN (c), NS-Pix2Pix (d), NG-DNN (e), NG-Pix2Pix (f).*

		PESQ						STOI					
	SNR	0	5	10	15	20	mean	0	5	10	15	20	mean
Airplane	(a)	1.34	1.63	2.02	2.47	3.00	2.09	0.64	0.74	0.82	**0.88**	**0.93**	0.80
	(b)	1.54	1.79	2.17	2.72	**3.26**	2.30	0.66	0.74	0.81	0.87	0.91	0.80
	(c)	1.65	1.94	2.30	2.73	3.16	2.36	**0.69**	**0.76**	**0.83**	**0.88**	0.92	**0.82**
	(d)	1.57	2.02	**2.51**	**2.91**	3.18	2.44	0.66	0.75	0.81	0.85	0.89	0.79
	(e)	1.65	1.94	2.29	2.70	3.14	2.35	**0.69**	**0.76**	0.82	0.87	0.91	0.81
	(f)	**1.67**	**2.07**	**2.51**	2.88	3.13	**2.45**	0.67	0.74	0.79	0.83	0.86	0.78
Babble	(a)	1.20	1.42	1.79	2.40	**3.13**	1.99	0.44	0.56	0.67	0.77	0.85	0.66
	(b)	1.14	1.31	1.61	2.07	2.65	1.76	0.43	0.56	0.66	0.74	0.81	0.64
	(c)	**1.25**	1.51	1.87	2.31	2.78	1.95	**0.50**	**0.63**	**0.72**	**0.79**	**0.86**	**0.70**
	(d)	1.20	1.48	1.98	2.52	2.93	2.02	0.46	0.59	0.71	0.78	0.83	0.67
	(e)	1.24	**1.52**	1.88	2.31	2.78	1.95	0.49	0.62	**0.72**	**0.79**	0.85	**0.70**
	(f)	1.20	1.49	**2.00**	**2.53**	2.93	**2.03**	0.46	0.60	0.71	0.77	0.82	0.67
Cantine	(a)	1.35	1.65	2.07	2.57	**3.30**	2.19	0.54	0.66	0.75	**0.83**	**0.90**	0.74
	(b)	1.38	1.68	2.12	2.67	3.23	2.22	0.55	0.66	0.74	0.82	0.87	0.73
	(c)	1.46	1.75	2.15	2.63	3.12	2.22	0.59	**0.69**	0.76	**0.83**	0.89	0.75
	(d)	1.45	1.84	2.38	**2.82**	3.13	2.32	0.58	0.68	0.75	0.80	0.85	0.73
	(e)	1.47	1.77	2.18	2.64	3.11	2.24	**0.60**	**0.69**	**0.77**	**0.83**	0.89	**0.76**
	(f)	**1.49**	**1.91**	**2.43**	2.81	3.08	**2.34**	0.59	**0.69**	0.75	0.80	0.84	0.73
Market	(a)	1.26	1.51	1.89	2.38	3.04	2.02	0.51	0.62	0.73	0.81	**0.88**	0.71
	(b)	1.24	1.45	1.76	2.22	2.79	1.89	0.51	0.62	0.71	0.79	0.85	0.70
	(c)	1.35	1.63	2.00	2.46	2.94	2.08	**0.56**	**0.67**	**0.75**	**0.82**	**0.88**	**0.73**
	(d)	**1.36**	1.71	2.21	**2.72**	**3.09**	**2.22**	0.55	0.66	0.74	0.80	0.85	0.72
	(e)	**1.36**	1.63	2.00	2.45	2.93	2.07	**0.56**	**0.67**	**0.75**	**0.82**	**0.88**	**0.73**
	(f)	1.35	**1.72**	**2.24**	2.68	3.02	2.20	**0.56**	**0.67**	0.74	0.79	0.83	0.72
White	(a)	1.15	1.31	1.60	2.01	2.57	1.73	0.50	0.61	0.72	0.81	**0.89**	0.71
	(b)	1.35	1.58	1.88	2.25	2.71	1.95	0.53	0.63	0.73	0.81	0.87	0.72
	(c)	**1.38**	1.66	2.00	2.39	2.88	2.06	**0.58**	**0.67**	**0.75**	**0.82**	0.88	**0.74**
	(d)	1.23	1.54	2.11	**2.74**	**3.14**	2.15	0.53	0.64	0.73	0.80	0.86	0.71
	(e)	1.35	1.63	1.96	2.29	2.65	1.98	0.57	0.66	0.74	0.81	0.88	0.73
	(f)	1.32	**1.69**	**2.22**	2.68	3.01	**2.19**	0.55	0.65	0.73	0.78	0.83	0.71

The ASV performances (Tests 2 and 3) are reported in Tables 2 and 3, where the results of the baseline systems are from [38]. For the clean speaker models, Pix2Pix front-ends generally outperform the baseline methods. One exception is seen for Babble noise, where the NG-DNN front-end gives an EER of 8.73%, marginally better than NS-Pix2Pix (8.76%). At low SNR, DNN-SE front-ends sometimes show better results than Pix2Pix, but overall our approach can be considered superior.

On the other hand, the performances of DNN-SE front-ends on multi-condition training are generally better, which presents a substantial improvement if compared with the clean speaker model situation. Our approach is generally better than STSA-MMSE, although the NS-Pix2Pix front-end shows lower per-

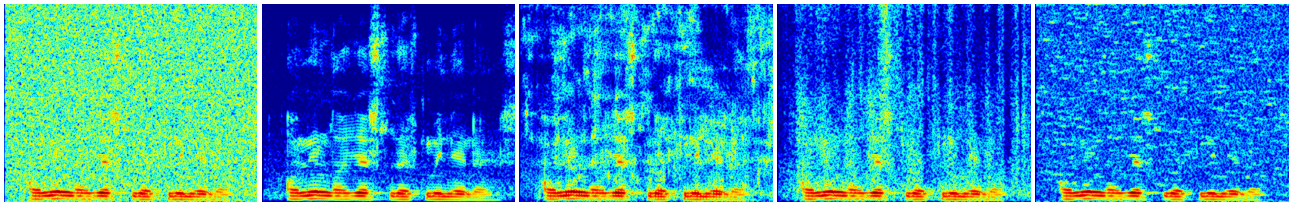

Figure 2: *From left to right: noisy spectrogram (White noise at 0 dB SNR); clean spectrogram; spectrogram of the signal enhanced with NG-Pix2Pix; spectrogram of the signal enhanced with NG-DNN; spectrogram of the signal enhanced with STSA-MMSE.*

Table 2: *ASV performance in terms of EER on clean speaker model*

	SNR	0	5	10	15	20	clean	mean
Airplane	No enhancement	21.09	15.99	13.61	11.66	9.18	6.99	13.08
	STSA-MMSE	17.69	12.58	8.17	6.53	6.27	5.80	9.51
	NS-DNN	16.99	10.55	7.48	6.99	6.15	6.12	9.05
	NS-Pix2Pix	17.19	8.84	**5.44**	5.05	**4.63**	**3.74**	7.48
	NG-DNN	15.99	8.99	6.12	6.12	5.58	5.67	8.08
	NG-Pix2Pix	**15.31**	**7.89**	5.58	**4.77**	4.76	5.44	**7.29**
Babble	No enhancement	19.05	14.63	11.69	11.04	9.18	6.99	12.10
	STSA-MMSE	29.04	20.40	12.59	7.82	6.29	5.80	13.66
	NS-DNN	17.01	10.54	7.82	6.46	6.12	5.78	8.96
	NS-Pix2Pix	18.83	11.22	7.62	**5.70**	5.10	**4.08**	8.76
	NG-DNN	**16.67**	**10.39**	**7.50**	6.34	5.78	5.67	**8.73**
	NG-Pix2Pix	21.05	13.64	8.50	5.97	**4.76**	5.44	9.90
Cantine	No enhancement	20.72	19.20	14.74	11.81	8.50	6.99	13.66
	STSA-MMSE	19.09	12.37	8.16	6.80	6.12	5.80	9.72
	NS-DNN	18.71	**8.58**	6.12	5.49	5.31	5.10	8.22
	NS-Pix2Pix	**17.33**	9.18	**5.44**	**5.10**	5.10	**4.16**	**7.72**
	NG-DNN	19.94	9.18	6.12	5.78	5.44	5.67	8.69
	NG-Pix2Pix	17.57	8.84	5.73	5.31	**4.76**	5.44	7.94
Market	No enhancement	29.40	20.07	15.00	11.96	8.93	6.99	15.39
	STSA-MMSE	25.51	17.35	11.90	8.28	7.35	5.80	12.70
	NS-DNN	21.43	**9.86**	**6.88**	6.46	5.78	5.92	9.39
	NS-Pix2Pix	**17.91**	10.33	7.14	**5.92**	5.17	**3.61**	**8.35**
	NG-DNN	21.77	10.59	7.48	6.22	5.76	5.67	9.58
	NG-Pix2Pix	19.58	11.22	7.48	6.12	**5.07**	5.44	9.15
White	No enhancement	45.90	43.20	34.61	26.28	16.91	6.99	28.98
	STSA-MMSE	30.95	21.17	13.95	10.20	8.50	5.80	15.10
	NS-DNN	39.46	20.75	9.86	7.82	6.12	6.02	15.01
	NS-Pix2Pix	40.48	28.23	12.45	7.86	6.46	6.46	16.99
	NG-DNN	40.14	21.77	10.88	8.16	6.80	5.67	15.57
	NG-Pix2Pix	**30.61**	**17.33**	**9.40**	**7.14**	**5.78**	**5.44**	**12.62**

Table 3: *ASV performance in terms of EER on multi-condition speaker model*

	SNR	0	5	10	15	20	clean	mean
Airplane	No enhancement	32.28	26.87	21.10	16.38	9.86	5.83	18.72
	STSA-MMSE	25.51	15.48	8.16	6.12	5.44	5.44	11.03
	NS-DNN	14.78	8.26	5.44	5.53	4.76	4.76	7.26
	NS-Pix2Pix	16.67	7.14	5.10	**4.03**	**3.78**	4.42	6.86
	NG-DNN	**11.38**	**6.12**	**4.78**	4.72	4.23	**4.00**	**5.87**
	NG-Pix2Pix	13.27	6.43	5.78	5.44	5.27	4.78	6.83
Babble	No enhancement	21.77	15.37	11.93	9.52	8.16	6.12	12.15
	STSA-MMSE	33.50	23.13	16.23	12.63	8.84	7.12	16.91
	NS-DNN	16.26	9.52	6.99	6.08	5.78	5.17	8.30
	NS-Pix2Pix	20.75	10.88	6.12	**4.76**	**4.08**	4.36	8.49
	NG-DNN	**16.00**	**9.18**	**5.44**	**4.76**	**4.08**	**4.00**	**7.19**
	NG-Pix2Pix	21.72	12.44	6.46	5.34	5.22	4.78	9.33
Cantine	No enhancement	24.11	17.22	12.93	10.88	9.18	7.48	13.63
	STSA-MMSE	19.05	12.59	8.21	6.91	6.12	6.32	9.87
	NS-DNN	12.93	5.91	**4.42**	4.25	4.27	**3.78**	5.93
	NS-Pix2Pix	14.29	6.87	4.76	**4.00**	**4.08**	4.76	6.46
	NG-DNN	**11.61**	**5.78**	5.10	4.57	**4.08**	4.00	**5.86**
	NG-Pix2Pix	14.10	7.48	5.44	5.44	5.27	4.78	7.08
Market	No enhancement	36.05	26.06	18.37	13.32	9.18	5.44	18.07
	STSA-MMSE	29.25	21.07	13.95	10.98	7.82	6.67	14.97
	NS-DNN	19.33	**8.16**	6.24	5.41	4.53	4.29	7.99
	NS-Pix2Pix	18.49	9.18	5.82	**4.42**	**3.74**	4.76	7.74
	NG-DNN	**18.37**	**8.16**	**5.78**	4.44	4.42	**4.00**	**7.53**
	NG-Pix2Pix	19.30	9.37	6.37	5.44	5.10	4.78	8.39
White	No enhancement	35.88	24.40	18.37	15.81	14.97	5.85	19.21
	STSA-MMSE	30.95	20.07	7.48	6.46	6.46	4.76	12.70
	NS-DNN	27.21	**9.52**	**6.12**	**5.02**	4.65	5.78	9.72
	NS-Pix2Pix	39.37	23.81	10.20	6.46	5.95	6.44	15.37
	NG-DNN	**26.19**	11.22	7.14	5.10	**4.08**	**4.00**	**9.62**
	NG-Pix2Pix	30.41	14.29	8.84	6.60	5.78	4.78	11.78

formance when it deals with white noise.

In general, Pix2Pix can be considered competitive with DNN-SE (better PESQ and EER on the clean speaker models, but worse STOI and EER for multi-condition training) and overall superior to STSA-MMSE.

Figure 2 shows the spectrograms of a noisy utterance (White noise at 0 dB SNR), together with its clean and enhanced versions with NG-Pix2Pix, NG-DNN, and STSA-MMSE. It is observed that the spectrogram enhanced by the cGAN approach preserves the structure of the original signal better than the other SE techniques, while at the same time more noises remain especially at high frequency regions, as compared with NG-DNN. The spectrogram enhanced by STSA-MMSE is snowy all over the places.

5. Conclusion

In this paper we investigated the use of conditional generative adversarial networks (cGANs) for speech enhancement. We adapted the Pix2Pix framework, intended to solve generic image-to-image translation problems, and evaluated the performance of this approach in terms of estimated speech perceptual quality and speech intelligibility, together with equal error rate of a Gaussian Mixture Model - Universal Background Model based speaker verification system. The results we obtained allow us to conclude that cGANs are a promising technique for speech denoising, being globally superior to the classical STSA-MMSE algorithm, and comparable to a DNN-SE algorithm.

Future work includes a more extensive evaluation of the framework in more critical SNR situations, and modifications aiming at making it specific for the task. For example, a model with G generating a small size output window from a fixed number of successive frames can be built as it is often done in deep neural networks for speech processing, and a specific perceptual loss to be added to the cGAN loss can be designed.

6. Acknowledgements

The authors would like to thank Hong Yu for providing data and speaker verification results for the baseline systems and Morten Kolbæk for his assistance and software used for the speaker verification and DNN speech enhancement baseline systems.

This work is partly supported by the Horizon 2020 OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, and the iSocioBot project, funded by the Danish Council for Independent Research - Technology and Production Sciences (#1335-00162).

7. References

[1] 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.

[2] M. Kolbœk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 305–311.

[3] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[4] 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*, vol. 139, no. 5, pp. 2604–2612, 2016.

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

[6] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[7] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

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

[9] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[10] 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.

[11] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

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

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

[14] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

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

[16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016.

[17] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *arXiv preprint arXiv:1701.05957*, 2017.

[18] S. Mobin and J. Bruna, "Voice conversion using convolutional neural networks," *arXiv preprint arXiv:1610.08927*, 2016.

[19] S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee, "Deep visual analogy-making," in *Advances in Neural Information Processing Systems*, 2015, pp. 1252–1260.

[20] O. Mogren, "C-rnn-gan: Continuous recurrent neural networks with adversarial training," *arXiv preprint arXiv:1611.09904*, 2016.

[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[22] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 318–335.

[23] 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*, 2016, pp. 2536–2544.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[25] Y.-C. Lin, "pix2pix-tensorflow," Github repository: https://github.com/yenchenlin/pix2pix-tensorflow, 2016, accessed: March 2017.

[26] 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 *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 749–752.

[27] ITU, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," Available: https://www.itu.int/rec/T-REC-P.862.2-200511-S/en, 2005, accessed: March 2017.

[28] 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.

[29] A. K. Sarkar and Z.-H. Tan, "Text dependent speaker verification using un-supervised hmm-ubm and temporal gmm-ubm," *Proceedings of INTERSPEECH (to appear)*, 2016.

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

[31] R. C. Hendriks, R. Heusdens, and J. Jensen, "Mmse based noise psd tracking with low complexity," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4266–4269.

[32] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[33] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[34] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[35] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[37] M. Falcone, B. Fauve, M. Cornacchia *et al.*, "Corpora collection," *OCTAVE (Objective Control of TAlker VErification), Deliverable 17*, 2016.

[38] H. Yu, Z.-H. Tan, Z. Ma, and J. Guo, "Adversarial network bottleneck features for noise robust speaker verification," *Proceedings of INTERSPEECH (to appear)*, 2017.