Speech Enhancement Using Augmented SSL CycleGAN

Branislav Popović
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
bpopovic@uns.ac.rs

Lidija Krstanović
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
lidijakrstanovic@uns.ac.rs

Marko Janev
Institute of Mathematics
Serbian Academy of Sciences and Arts
Belgrade, Serbia
markojan@uns.ac.rs

Siniša Suzić
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
sinisa.suzic@uns.ac.rs

Tijana Nosek
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
tijana.nosek@uns.ac.rs

Jovan Galić
Faculty of Electrical Engineering
University of Banja Luka
Banja Luka, Bosnia and Herzegovina
jovan.galic@etf.unibl.org

***Abstract*—The purpose of a single-channel speech enhancement is to attenuate the noise component of noisy speech to increase the intelligibility and the perceived quality of the speech component. One such approach uses deep neural networks to transform noisy speech features into clean speech by minimizing the mean squared errors between the degraded and the clean features using paired datasets. Most recently, an unpaired datasets approach, CycleGAN speech enhancement, was proposed, obtaining state-of-the-art results, regardless there was no supervision during the actual training. Also, only a small amount of noisy speech data is usually accessible in comparison to clean speech. Therefore, in this paper, an augmented semi-supervised CycleGAN speech enhancement algorithm is proposed, where only a small percentage of the training database contains the actual paired data. This, as a consequence, prevents overfitting of the discriminator corresponding to the scarce noised speech domain during the initial training stages and also augments the discriminator by periodically adding clean speech samples transformed by the inverse network into the pool of the discriminator of the scarce noisy speech domain. Significantly better results in the means of several standard measures are obtained using the proposed augmented semi-supervised method in comparison to the baseline CycleGAN speech enhancement approach operating on a reduced noisy speech domain.**

***Index Terms*—augmented CycleGAN, speech enhancement, semi-supervised learning**

I. Introduction

Speech enhancement of the perturbed, i.e., noisy speech, is a very important front-end pre-processing stage component, aimed to improve the performance of automatic speech recognition (ASR) [1]- [5] and speaker recognition systems under noisy conditions [6], [7]. This enhancement is commonly used in mobile speech communications, voice assistants, hearing aids, smart surroundings, etc.

This research was supported by the Science Fund of the Republic of Serbia, #6524560, AI-S-ADAPT, and by the Serbian Ministry of Education, Science and Technological Development through the project no. $45103-68/2020-14/200156$: "Innovative Scientific and Artistic Research from the Faculty of Technical Sciences Activity Domain".

Deep learning neural network based methods have recently achieved state-of-the-art results in a single-channel speech enhancement. Initial attempts based on mask learning approach are presented in [8], in which a feature estimation algorithm was proposed to estimate the ideal ratio mask or the ideal binary mask on noisy input features in the Mel frequency domain using a nonlinear mapping represented by deep neural networks (DNNs) and a set of time-frequency unit level features. Feature mapping approaches are reported in [9]- [16], where the DNN mapping serves as a nonlinear regression function. These approaches are based on the assumption that the scale of the masked signal is the same as the scale of the clean target and the noise is strictly additive. Nevertheless, overfitting often occurs using the mentioned mean squared error (MSE) approach and the network does not generalize well to the unseen data, especially when the training dataset is small. Recently, an approach inspired by cycle-consistent adversarial networks [17] has been proposed. Apart from direct mapping F from noisy to clean domain, an additional mapping G from clean to noisy data is introduced. Both of these mappings are inverse to each other up "to a certain extent", which is enforced by adding additional losses $E_{x\sim p_X(x)}\|x-G(F(x))\|$ and $E_{y\sim p_Y(y)}\|y-F(G(y))\|$, where p_X and p_Y are distributions of noisy and clean data, respectively.

Although cycle-consistency can effectively regularize the structured data by enforcing bijectivity of direct and inverse network mappings, in many situations it is very hard, if not impossible, to obtain a sufficient amount of paired data to train the networks implementing those mappings. A framework for estimating generative models via an adversarial process is introduced in [18]. The concept achieved great success in the areas of image generation (see [19]- [23]) as well as image-to-image translation (see [17], [24]- [26]), and it has also been applied in the areas of speech enhancement [27]- [30] and perturbation invariant ASR [31], [32]. Most

recently, CycleGAN networks in which the adversarial loss is combined with the cycle-consistency loss obtained state-of-the art results in an unpaired image-to-image translation (see [26]) and the method is also applied with the similar state-of-the-art efficiency in both speech enhancement [33], [34] and non-parallel voice conversion tasks [35]- [37]. In this work, we introduce additional modifications of the CycleGAN concept, which we refer to as "Aug SSL CycleGAN", to deal with the problem when the noisy speech domain is scarce, i.e., the domain contains much fewer training samples than the clean speech domain, which is a very common situation in speech enhancement. For example, in the ASR task, large databases of clean speech are mostly available for the training of ASR systems, while there are much less noisy speech data available, especially as those should be separately collected for each specific task. We invoke two different strategies to cope with the mentioned issues. Firstly, a semi-supervised learning (SSL) strategy is applied using only a small percentage of paired, i.e. labelled training data samples which are combined into $\|\cdot\|_1$ norm loss component to prevent overfitting of the discriminator related to the scarce noisy speech domain. Secondly, after a number of initial iterations, data augmentation is introduced by periodically adding samples generated by the inverse mapping from the full (clean) to the scarce (noisy) domain. Experimental results on speech samples containing different amounts of artificially added Gaussian and various other types of noises show that the proposed Aug SSL CycleGAN speech enhancement method obtained significantly better results in comparison with the baseline CycleGAN approach.

II. Generative Adversarial Networks

In [18], a method of learning the ground truth data distribution in a nonparametric way is proposed in the form of Generative Adversarial Networks. The task is as follows: The discriminator network D is trying to discriminate between the samples generated by the generator network G and the ground truth observations. The generator models the ground truth data distribution by learning how to confuse the discriminator. Both of them are competing with each other to reach the equilibrium expressed by the mini-max loss of the training procedure. Optimization problem is therefore given by:

$$\min_G \max_D E_{x\sim p(x)} \ln[D(x)] + E_{z\sim p(z)} \ln[(1 - D(G(z)))] \quad (1)$$

where $p(x)$ represents the ground truth data distribution, while the latent variable z is sampled by the distribution $p(z)$.

III. CycleGAN Speech Enhancement

CycleGAN Speech Enhancement algorithm [33]- [34] uses two GANs in the opposite directions and combine adversarial and cycle consistency losses in order to train the mentioned networks in the task of transition from noisy to clean speech, in the case of unparalleled data, i.e., in the absence of paired training samples. Let us denote the noisy speech domain as X and the clean speech domain as Y, a direct mapping from noisy to clean speech domain by $G_{X\to Y}$, and an inverse mapping from the clean speech to the noisy speech domain by $G_{Y\to X}$ (both of them implemented in GAN). Let us also denote the discriminator corresponding to the noisy speech domain by D_X and the discriminator corresponding to the clean speech domain by D_Y, respectively. Cycle consistency loss forces $G_{X\to Y}$ and $G_{Y\to X}$ to be inverse to each other "as close as possible" i.e., $x \approx G_{Y\to X}(G_{X\to Y}(x))$, for all $x \in X$, as well as $y \approx G_{X\to Y}(G_{Y\to X}(y))$, for all $y \in Y$. Thus, $G_{X\to Y}$ will be capt "close to bijectivity", causing only small regions of the noisy domain X to map to the same clean speech instance $y \in Y$, by the mapping $G_{X\to Y}$. Therefore, we obtain the following adversarial objectives:

$$\begin{aligned}
&\mathcal{L}_{adv}(G_{X\to Y}, D_Y) = \\
&E_{y\sim p_Y(y)}[\ln D_Y(y)] + \\
&E_{x\sim p_X(x)}[\ln(1 - D_Y(G_{X\to Y}(x)))], \\
&\mathcal{L}_{adv}(G_{Y\to X}, D_X) = \\
&E_{x\sim p_X(x)}[\ln D_X(x)] + \\
&E_{y\sim p_Y(y)}[\ln(1 - D_X(G_{Y\to X}(y)))]
\end{aligned} \quad (2)$$

a cycle consistency objective:

$$\begin{aligned}
&\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) = \\
&E_{x\sim p_X(x)}[\|G_{Y\to X}(G_{X\to Y}(x)) - x\|] + \\
&E_{y\sim p_Y(y)}[\|G_{X\to Y}(G_{Y\to X}(y)) - y\|]
\end{aligned} \quad (3)$$

and the identity mapping loss given by:

$$\begin{aligned}
&\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) = \\
&E_{x\sim p_Y(y)}[\|G_{X\to Y}(y) - y\|] + \\
&E_{x\sim p_X(x)}[\|G_{Y\to X}(x) - x\|]
\end{aligned} \quad (4)$$

making the full objective given by:

$$\begin{aligned}
&\mathcal{L}(G_{X\to Y}, G_{Y\to X}, D_X, D_Y) = \\
&\mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \\
&\mathcal{L}_{adv}(G_{Y\to X}, D_X) + \\
&\lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) + \\
&\lambda_{id}\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X})
\end{aligned} \quad (5)$$

to be optimized, where $\lambda_{cyc} > 0$ and $\lambda_{id} > 0$ are the mixing coefficients.

IV. The Proposed Approach: Aug SSL CycleGAN Speech Enhancement

In order to deal with the problem when the noisy speech domain is scarce, two different strategies are invoked. We propose a semi-supervised learning (SSL) strategy based on a simple $\|\cdot\|_1$ norm error between the transformed noisy speech and the ground truth clean speech samples as well as between the clean speech and the ground truth noisy speech samples, i.e., the error is calculated in both directions, on a predefined amount of paired (labelled) training samples $\{(x_i, y_i)|i = 1, \ldots, m\} \subset X \times Y$ available, where m is the number of those pairs, $x_i \in X$ are the noisy speech samples,

while $y_i \in Y$ are their clean speech counterparts. Thus, by applying the previous method during the initial training stages, we prevent discriminator to overfit, which would otherwise be inevitable, due to the limited amount of data samples belonging to the scarce noisy speech domain. This additional SSL part of the loss function is then given by:

$$\begin{aligned}&\mathcal{L}_{SSL}(G_{X\to Y}, G_{Y\to X}) = \\ &\frac{1}{m}\sum_{i=1}^{m}[\|G_{X\to Y}(x_i) - y_i\|_1 + \\ &\|G_{Y\to X}(y_i) - x_i\|_1]\end{aligned} \tag{6}$$

making the full objective given by

$$\begin{aligned}&\mathcal{L}(G_{X\to Y}, G_{Y\to X}, D_X, D_Y) = \\ &\mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \\ &\mathcal{L}_{adv}(G_{Y\to X}, D_X) + \\ &\lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) + \\ &\lambda_{id}\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) + \\ &\lambda_{SSL}\mathcal{L}_{SSL}(G_{X\to Y}, G_{Y\to X})\end{aligned} \tag{7}$$

where $\lambda_{SSL} > 0$ is the mixing coefficient.

Secondly, we augment the discriminator corresponding to the scarce noisy speech domain by adding the samples generated by the inverse network mapping from the full, i.e., the clean speech domain to the scarce, i.e., the noisy speech domain, after a number of initial iterations, where D_X is forced not to overfit by applying the above mentioned SSL strategy. Namely, when the discriminator as well as the generator corresponding to the scarce, i.e, the noisy speech domain are sufficiently pre-trained using a combination of adversarial and the SSL loss, we perform the following: Samples generated by the pre-trained generator $G_{Y\to X}$ are added to the pool of data corresponding to the scarce, i.e., the noisy speech domain X, thus modifying its statistics as $p(X) \to \hat{p}(X)$. More precisely, in every k-th training iteration, additional samples generated by the network $G^{(k)}_{Y\to X}$ which transforms clean speech samples to the scarce, i.e., noisy speech domain in an inverse mapping procedure, are added to the pool of the discriminator of the noisy speech domain D_X, where (k) denotes the state of parameters at k-th iteration of training. Thus, in every k-th iteration, after a predefined number of epochs, the discriminator D_X of the noisy speech domain is trained using the data pool augmented in the previously described manner. In our case, the described procedure is applied after the first 4000 epochs, when the discriminator is sufficiently trained.

V. Network Architecture

We use an architecture proposed in [35], which represents a modification to the standard CycleGAN architecture presented in [26]. Namely, this architecture includes a gated CNN invoked in [38], which not only allows parallelization over sequential data, but also achieves state-of-the-art in speech modelling, as well as identity mapping loss invoked in [39]. The parameters of this network are set in the following way: The number of epochs was set to 5000 (so there was only a little or no overfitting), mini batch size to 1, generator learning rate (for both generators) to $\eta_G = 0.0002$, generator learning rate decay to $\nu_G = 0.0002/2000000$, discriminator learning rate (for both discriminators) to $\eta_D = 0.0001$, discriminator learning rate decay to $\nu_D = 0.0002/2000000$, $\lambda_{cyc} = \lambda_{SSL} = 10$, and $\lambda_{id} = 5$. 24 Mel-cepstral coefficients, logarithmic fundamental frequency and aperiodicities are extracted every 5 ms from a randomly chosen fixed-length segment of 128 frames. One-dimensional CNN is used as generator to capture the relationship among features while preserving the temporal structure.

VI. Experimental Results

In this section, we present experimental results obtained on a commercially available speech database, recorded at AlfaNum Speech Technologies in Novi Sad, containing 200 clean and 200 artificially created noisy samples in the training dataset (10 male and 10 female speakers, 10 audio files per each), as well as 50 clean samples and the same number of their noisy counterparts (5 male and 5 female speakers, 5 samples each). The noise is combination of an additive stationary (Gaussian) noise and various kinds of non-stationary components, such as traffic and office noise, crackling, creaking, etc. Samples are recorded as mono, PCM signals, at 16000 Hz sampling rate, 16 bits per sample.

Perceptual Evaluation of Speech Quality (PESQ) is a family of standards used for the objective voice quality assessments by phone manufacturers, network equipment vendors and telecom operators (ITU-T P.862 standard recommendation). Two different measures of speech signal quality are obtained using these standards, PEQMOS and MOSLQO.

The second standard used in this paper is a Virtual Speech Quality Objective Listener (ViSQOL). It is a signal-based, full-reference, intrusive metric that models human speech quality perception using a spectro-temporal measure of similarity between referent and degraded speech signal. The metric is particularly designed for quality assessments associated with Voice over IP (VoIP) transmissions. Using ViSQOL, other two measures are obtained, VISQOL and NSIM.

In Tables I and II, average results over 50 test samples on the noisy speech↔clean speech transformation task are presented using PEQMOS and MOSLQO measures, respectively. The measures are evaluated by comparison of the transformed degraded, i.e., noisy speech signals to corresponding clean speech samples, for the proposed Aug SSL CycleGAN speech enhancement in comparison to the baseline CycleGAN speech enhancement algorithm, for various percentages of the paired noisy-clean samples given for the scarce domain. SSL denotes that only SSL strategy is applied on the control part of labelled training samples included in the training of the CycleGAN. Aug denotes that only augmentation of the pool of the discriminator corresponding to the scarce, i.e., the noisy speech domain is applied. SSL+Aug denotes that both SSL and Aug strategies are applied. The proposed method obtained better

results in comparison to the baseline CycleGAN algorithm on the scarce, noisy speech domain, for both of these measures.

In Tables III and IV, VISQOL and NSIM measures are presented in the same manner as above. For both of these measures, the Aug SSL CycleGAN speech enhancement method showed significantly better results in comparison to the baseline CycleGAN method. Improvements were generally consistent with the increase in the number of paired speech samples in the SSL component of the training, as well as with adding the augmentation component of the training (a more significant improvement of the results is achieved by combining these two strategies).

In Fig. 1, spectrograms are presented for one of the signals transformed using the above-mentioned algorithms. In Fig. 2, spectrograms are presented for the selected part of the signal which contains only noise. Significant noise reduction can be observed in both of these figures while preserving the vocal component. The same conclusion can be drawn by subjective comparison (listening tests) of different audio signals generated using these algorithms. 10 students from the Faculty of Technical Sciences in Novi Sad provided their subjective evaluations (scale 1 to 5) of the original signals (average score of 2.7) and their transformed counterparts (3.4 for the baseline and 3.9 for the SSL+Aug version).

TABLE I
PEQMOS: NOISY SPEECH↔CLEAN SPEECH TASK

[%]	CycleGAN	SSL	Aug	SSL+Aug
25%	-	0.831	0.832	0.837
50%	-	0.864	0.842	0.857
100%	0.828	0.853	0.850	0.867

TABLE II
MOSLQO: NOISY SPEECH↔CLEAN SPEECH TASK

[%]	CycleGAN	SSL	Aug	SSL+Aug
25%	-	1.179	1.171	1.178
50%	-	1.184	1.186	1.191
100%	1.178	1.211	1.232	1.238

TABLE III
VISQOL: NOISY SPEECH↔CLEAN SPEECH TASK

[%]	CycleGAN	SSL	Aug	SSL+Aug
25%	-	1.384	1.399	1.392
50%	-	1.409	1.410	1.425
100%	1.369	1.436	1.391	1.452

TABLE IV
NSIM: NOISY SPEECH↔CLEAN SPEECH TASK

[%]	CycleGAN	SSL	Aug	SSL+Aug
25%	-	0.572	0.573	0.579
50%	-	0.572	0.575	0.581
100%	0.569	0.576	0.575	0.585

VII. CONCLUSION

In this paper, a novel augmented semi-supervised CycleGAN speech enhancement approach is presented, using various strategies to improve transformation results by applying only a small percentage of labelled train data, preventing at the same time overfitting of the discriminator. The results were evaluated using various objective measures, by observing the spectrograms, as well as subjectively. Significantly better results are obtained in comparison to the baseline algorithm. However, both of these algorithms result in a noticeable signal reduction around non-stationary noise components, which could be the focus of our future work. The proposed methodology can also be applied to various other tasks, such as e.g., speech style transformation (using appropriate alignment methodologies).

REFERENCES

[1] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, 2012.

[2] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," INTERSPEECH, 2012.

[3] T. Sainath, B. Kingsbury, B. Ramabhadran et al., "Making deep belief networks effective for large vocabulary continuous speech recognition," ASRU, pp. 30–35, 2011.

[4] L. Deng, J. Li, J. T. Huang et al., "Recent advances in deep learning for speech research at Microsoft," ICASSP, pp. 8604–8608, 2013.

[5] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," IEEE/CAA J. Autom. Sin., vol. 4, no. 3, pp. 396–409, 2017.

[6] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 22, no. 4, pp. 745–777, April 2014.

[7] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, Robust Automatic Speech Recognition: A Bridge to Practical Applications. Academic Press, 2015.

[8] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," ICASSP, pp. 7092–7096, 2013.

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

[10] F. Weninger, H. Erdogan, S. Watanabe et al., "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," LVA/ICA, pp. 91–99, 2015.

[11] 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 Lang. Process., vol. 23, no. 1, pp. 7–19, 2015.

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

[13] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," INTERSPEECH, 2012.

[14] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," ICASSP, pp. 1759–1763, 2014.

[15] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," ICASSP, pp. 3709–3713, 2014.

[16] Z. Chen, Y. Huang, J. Li, and Y. Gong, "Improving mask learning based speech enhancement system with restoration layers and residual connection," INTERSPEECH, pp. 3632–3636, 2017.

[17] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," CVPR, pp. 1125–1134, 2017.

[18] Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," NIPS, pp. 2672–2680, 2014.

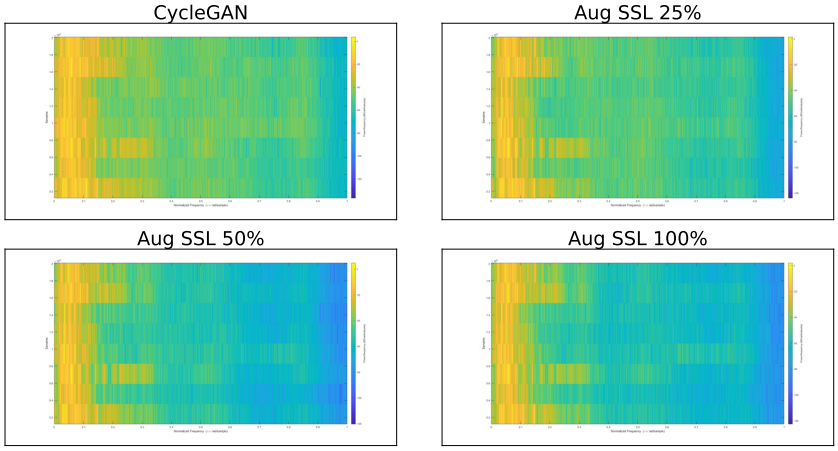

Fig. 1. Spectrograms (full signal): CycleGAN vs. Aug SSL 25/50/100%

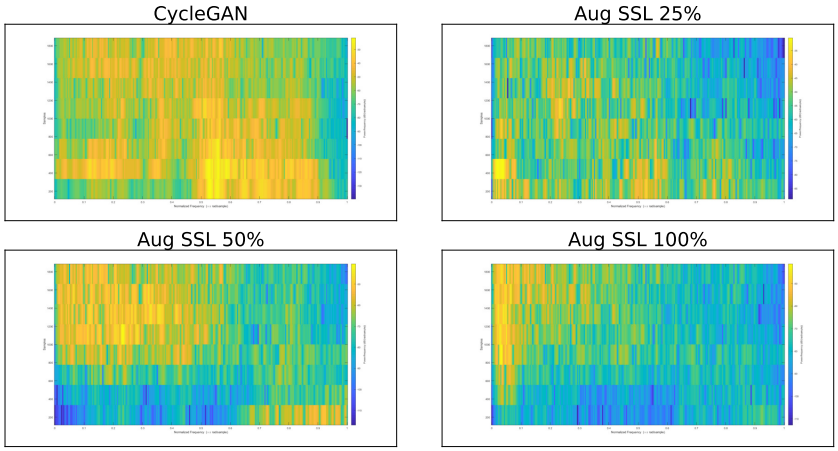

Fig. 2. Spectrograms (noise): CycleGAN vs. Aug SSL 25/50/100%

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

[20] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," NIPS, pp. 1486–1494, 2015.

[21] T. Park, M. Y. Liu, T. C. Wang, and J. Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," CVPR, pp. 2337–2346, June 2019.

[22] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, Sean: "Image synthesis with semantic region-adaptive normalization," CVPR, pp. 5104–5113, June 2020.

[23] C. H. Lee, Z. Liu, L. Wu, and P. Luo, Maskgan: "Towards diverse and interactive facial image manipulation," CVPR, pp. 5549–5558, 2020.

[24] C. Wang, H. Zheng, Z. Yu, Z. Zheng, Z. Gu, and B. Zheng, "Discriminative region proposal adversarial networks for high-quality image-to-image translation," ECCV, pp. 770–785, September 2018.

[25] B. AlBahar and J.-B. Huang, "Guided image-to-image translation with bi-directional feature transformation," ICCV, pp. 9016–9025, October 2019.

[26] Zhu, J., Park, T., Isola, P., Efros, A.A.: "Unpaired image-to-image translation using cycle-consistent adversarial networks," ICCV, pp. 2242–2251, 2017.

[27] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," INTERSPEECH, pp. 3642–3646, 2017.

[28] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," arXiv:1711.05747 [cs.SD], 2017.

[29] M. Mimura, S. Sakai, and T. Kawahara, "Cross-domain speech recognition using nonparallel corpora with cycle-consistent adversarial networks," ASRU, pp. 134–140, 2017.

[30] Z. Meng, J. Li, Y. Gong, and B.-H. F. Juang, "Adversarial feature-mapping for speech enhancement," INTERSPEECH, pp. 3259–3263, 2018.

[31] G. Saon, G. Kurata, T. Sercu et al., "English conversational telephone speech recognition by humans and machines," INTERSPEECH, pp. 132–136, 2017.

[32] Z. Meng, J. Li, Z. Chen et al., "Speaker-invariant training via adversarial learning," ICASSP, pp. 5969-5973, 2018.

[33] Z. Meng, J. Li, Y. Gong, B.H. Juang, "Cycle-consistent speech enhancement," INTERSPEECH, pp. 1165–1169, 2018.

[34] S. H. Dumpala, R. Chakraborty, S. K. Kopparapu, I. Sheikh, "Improving ASR robustness to perturbed speech using cycle-consistent generative adversarial networks," ICASSP, pp. 5726–5730, 2019.

[35] T. Kaneko, H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," EUSIPCO, pp. 2114–2118, 2018.

[36] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion," ICASSP, pp. 6820–6824, 2019.

[37] T. Kaneko, H. Kameoka, K. Tanaka, N. Hojo, "CycleGAN-VC3: Examining and improving CycleGAN-VCs for Mel-spectrogram conversion," INTERSPEECH, pp. 2017-2021, 2020.

[38] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," ICML, pp. 933–941, 2017.

[39] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," ICLR, 2017.