10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Novel Inception-GAN for Whisper-to-Normal Speech Conversion

Maitreya Patel, Mihir Parmar, Savan Doshi, Nirmesh Shah and Hemant A. Patil

Speech Research Lab, DA-IICT, Gandhinagar-382007, India.
{maitreya_patel, mihir_parmar, savan_doshi, nirmesh88_shah, hemant_patil}@daiict.ac.in

Abstract

Recently, Convolutional Neural Networks (CNN)-based Generative Adversarial Networks (GANs) are used for Whisper-to-Normal Speech (i.e., WHSP2SPCH) conversion task. These CNN-based GANs are significantly difficult to train in terms of computational complexity. Goal of the generator in GAN is to map the features of the whispered speech to that of the normal speech efficiently. To improve the performance, we need to either tune the cost functions by changing hyperparameters associated with it or to make the generator more complex by adding more layers to the model. However, more complex architectures are prone to overfitting. Both solutions are time-consuming and computationally expensive. Hence, in this paper, we propose Inception-based GAN architecture (i.e., Inception-GAN). Our proposed architecture is quite stable and computationally less expensive while training. The proposed Inception-GAN outperforms existing CNN-based GAN architectures (CNN-GAN). Objective and subjective results are carried out using the proposed architectures on statistically meaningful whispered TIMIT (wTIMIT) corpus. For a speaker-specific evaluations, Inception-GAN shows 8.9% and 6.2% better perfomance objectively compared to the CNN-based GAN for male and female speaker, respectively.

Index Terms: Whisper, CNN, GAN, Inception, Inception-GAN.

1. Introduction

In recent years, Deep Learning (DL) has shown its significant performance in the domain of Voice Conversion (VC) and its various applications. In particular, Whispered-to-Normal Speech (i.e., WHSP2SPCH) conversion is one of the applications of VC. However, there are still many barriers in this field that exists from speech processing and machine learning (ML) perspective in WHSP2SPCH conversion problem [1]. WHSP2SPCH conversion has many applications, such as private communication in public places, communication in silent places like library, hospital, etc., whispered wake-up word recognition in Intelligent Personal Assistants (IPAs), medical-domain applications, and many others [1]. Accidentally, vocal folds [2, 3], larynx [4, 5] or other articulatory parts of people related to speech production have been affected, and some of these people can converse only by whispering. Loosing the natural way of speaking ability make these peoples' life miserable. Hence, WHSP2SPCH conversion is also important for these impaired people for their daily communication. For this, we need efficient and less expensive WHSP2SPCH conversion systems. In this paper, we proposed a Inception-GAN which shows more stable and promising results compared to the CNN-GAN [6].

Whispered and normal speeches are significantly different from speech production-perception perspective [1, 7–9]. The coupling between the trachea increases, that results in spectral differences between the whispered and normal speeches. Hence, phone duration, energy distribution across phone classes, formant locations, and the spectral tilt will also be affected [1, 8, 10]. Speech is generated due to the periodic vibrations of the vocal folds, while whispered speech has almost no vibrations of vocal folds [1, 11]. This period of vocal fold vibrations in speech is called fundamental or pitch frequency (F_0). These periodic vibrations are responsible for voiced sounds in speech. The less or no vibrations of vocal folds in whispered speech leads to unvoiced sounds. Hence, whispered speech is completely aperiodic or unvoiced. To generate normal speech, we need F_0 which is absent in whispered speech. Even though F_0 is missing in the whispered speech, it has been found that pitch is encapsulated in a tangled way in the whispered speech [12–15]. Hence, prediction of F_0 is one of the most strenuous and formidable task for WHSP2SPCH conversion task.

Right after the GAN was invented [16], many GAN-based architectures actively used for VC [7, 17–22]. Recently, Cycle-consistent Adversarial Network (CycleGAN) gives the state-of-the-art results in VC [23–26]. While other variants of GANs, such as MMSE DiscoGAN proposed in [27] is also shown significant results for WHSP2SPCH conversion task, in particular, for prediction of F_0. Most of these models are implemented using Deep Neural Networks (DNN)-based architecture for WHSP2SPCH conversion. While in VC, most of the architectures are based on CNN. Contradictory, mapping function based on traditional CNNs are relatively unstable and hard to train compared to the training of advanced GAN-based architectures, such as CycleGAN. The main issue that needs to be addressed is traditional GAN-based architectures (such as, Vanilla GAN, Wasserstein GAN, CNN-GAN, etc.) shows the problem of mode collapse more often and never converge since generator is not able to learn the mapping function between the features of whispered and normal speeches. Before generator learns the mapping, discriminator efficiently learns to discriminate between original and the converted speech features. To solve this problem, we can tune either the relative importance of the loss functions while training to make the generator learn more faster or add more CNN layers to make generator more efficient as compared to the discriminator. However, these solutions are immensely costly in terms of computation, and time constraints. Hence, in this paper, we propose Inception-GAN which uses Inception network into both generator, and discriminator to overcome these limitations.

Experimental results show that Inception-GAN is more stable (i.e., it is not hyperparameter sensitive) compared to the CNN-GAN, and it is able to map at different sparse-level at the same time. Alongside of these advantages of Inception-GAN, it has less computational complexity compared to CNN-based GAN. In brief, Inception-GAN outperforms CNN-GAN by keeping training simple, stable, and less complex. Furthermore, we provide speaker-specific quantitative, and qualitative results for Inception-GAN and CNN-GAN.

10.21437/SSW.2019-16

2. Proposed Architectures

In this Section, we briefly summarize the key idea of inception module proposed in [28, 29], and detailed analysis of our proposed architecture, namely, Inception-GAN.

2.1. Inception module

The simple method to increase the performance of any architectures is to increase the length of it (i.e., depth of the network) by adding more number of layers, and increase its width(i.e., bigger sized convolutions) as well [28]. However, this leads to the increase in computational complexity, and the probability of overfitting. Hence, there is a need of a method which can use sparsity locally, and can be replaced with CNN. In this paper, we achieved this by using the Inception architecture [29].

The idea of Inception is used in CNN as a building block in an intrinsic way. In this paper, we have used convolution layers parallel with filter size of 1×1, 3×3, and 5×5. In addition, we have used 3×3 max-pooling layers. Generally, max-pooling leads to information loss. However, we are extracting different features at the same time. Hence, the idea of using max-pooling layer is to extract more dominated features from the input. In this paper, we assume that the output of different filters that we have used parallelly may end up with the feature clusters centered at the specific input features. To address this issue, 1×1 convolution is used in the inception model. Mainly, the reasons of using 1×1 convolutions are that it will cluster outputs of previous layer with high correlations [28]. However, there will be some regions in input of previous layer, where features are not centered. Hence, we have used 3×3, and 5×5 convolutions. Therefore, we can say that inception uses the filter-level sparsity. The naive version of proposed Inception module is shown in figure 1.

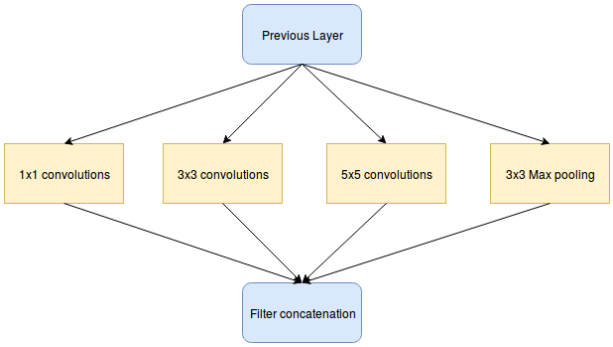

Figure 1: *Schematic representation of naive version of Inception module. After [28].*

Inception modules will be stacked upon each others just like CNN. Features will be filtered with good abstraction by upper Inception modules, and details of these abstractions will be increased layer-by-layer. Hence, as we move to the higher layers, the ratio of the number of 3×3, and 5×5 convolutions should increase. However, current Inception module does not solve the issue of computational complexity. The reason for this problem is that the computational complexity will be the sum of individual complexities of 1×1, 3×3, and 5×5 since we want the number of filters at the output layer same as the number of filters at input layer. To overcome this, one more convolution layer is used inside the Inception. Moreover, 1×1 convolutions are used to compute reductions before the expensive 3×3, and 5×5 convolutions (as shown in figure 2). In

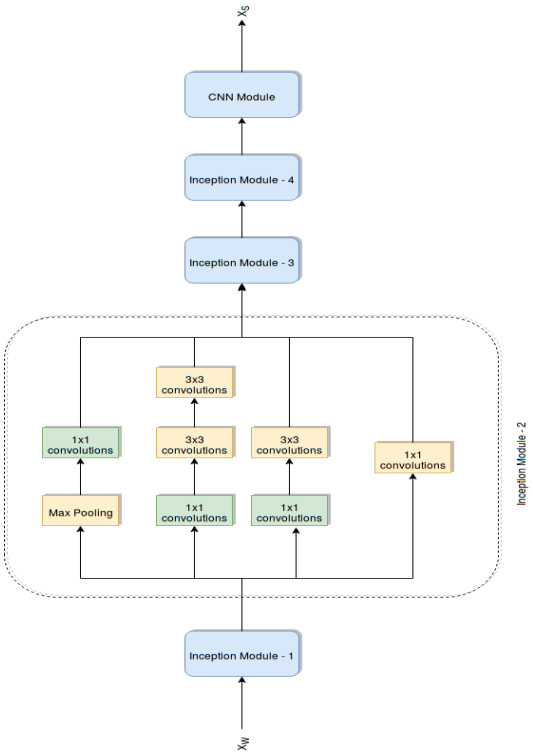

Figure 2: *Schematic representation of proposed Generator which uses Inception module for WHSP2SPCH conversion.*

addition, 5×5 convolutions are more expensive in terms of computations compared to the 3×3 convolutions. Assuming that there are n number of filters, computational complexity of 5×5 convolutions will be $25/9 = 2.78$ times more expensive than 3×3 convolutions. However, 5×5 convolutions cannot be just replaced with 3×3 convolutions since 5×5 convolutions captures spread-out features in a better way as compared to the 3×3 convolutions. Replacing 5×5 convolutions with the 3×3 leads to loss of linguistic information in converted normal speech. We know that the basic aim of WHSP2SPCH is to retain linguistic information while conversion. However, spatial aggregation can be done over lower-dimensional embedding without loosing the linguistic information [29]. One can reduce the dimension of the input representation (e.g., 3×3) before the spatial aggregation (e.g., 5×5) without expecting serious adverse effects. Hence, we have replaced 5×5 convolutions with two consecutive 3×3 convolutions, and reduced the computational complexity by $[25 - (9 + 9)]/25$, 0.28, times without loosing information compared to the Inception module (as shown in figure 2).

As discussed above, due to the less computational complexity of Inception module compared to the traditional CNN, one can use it to create larger architectures by increasing the *depth* as well as *width* of the network.

2.2. Proposed Inception-GAN

In Inception-GAN, generators and discriminators are made by stacking Inception modules on top of each other. Figure 2 shows how the Inception module is incorporated into the generator. This generator is made of four updated Inception modules, and a single CNN layer. Similarly, we have used Inception module

Table 1: *Proposed Architectural Details of Generator in Inception-GAN*

Module	**Patch size/ Stride/ Padding**	**# 1x1**	**(1) #3x3 reduce**	**(1) #3x3**	**(2) #3x3 reduce**	**(2) #3x3**	**(2) #3x3**	**Max Pool**
inception	-	32	50	64	16	32	32	16
inception	-	64	64	128	32	64	64	32
inception	-	64	64	128	32	64	64	32
inception	-	32	128	64	64	32	32	16
cnn	3 x 3/1/1	-	-	-	-	-	-	-

Table 2: *Proposed Architectural Details of Discriminator in Inception-GAN*

Module	**Patch size/ Stride/ Padding**	**Output Neurons**	**# 1x1**	**(1) #3x3 reduce**	**(1) #3x3**	**(2) #3x3 reduce**	**(2) #3x3**	**(2) #3x3**	**Max Pool**
convolution	7x7/2/3	-	-	-	-	-	-	-	-
inception	-	-	32	50	64	16	32	32	16
avg pool	(25, 2)/(5, 1)	-	-	-	-	-	-	-	-
inception	-	-	64	64	128	32	64	64	32
avg pool	3/2/1	-	-	-	-	-	-	-	-
inception	-	-	128	128	256	32	64	64	48
avg pool	(10, 2)/(6, 6)	-	-	-	-	-	-	-	-
linear	-	1028	-	-	-	-	-	-	-
dropout (50%)	-	-	-	-	-	-	-	-	-
linear	-	1	-	-	-	-	-	-	-

for discriminator also. Table 1 and Table 2 shows the details of the architecture for the generator, and discriminator of proposed Inception-GAN, respectively. In tables, "$\#3 \times 3$ reduce" shows the number of 1×1 filters that are used for reduction before expensive 3×3 convolution.

After the deployment of generator with real-life data, the size of input feature matrix will change for each input wave file (more details about the features is in Section 3). Hence, dimensions of input feature matrix varies for each input wave file while testing of generator in real-life. Moreover, when we have odd number of input frames, size of input feature matrix before downsampling, and size of output feature matrix after up sampling will be different (i.e., output frames will be always ± 1 of the number of input frames) which leads to problem of target dimensions mismatching while evaluations. To overcome all these limitations, we developed the generator such that it never downsamples the input feature matrix. After the Inception module in generator, we have used single convolution operation to cluster stacked output of previous Inception module. In the case of discriminator, we have used average pooling after every inception module to decrease the size of the input feature matrix. After four modules of Inception, we have used two fully-connected neural networks with a dropout to slow down the learning speed of discriminator at some extent.

2.3. Training Methodology

To learn the mapping function between features of the whispered speech (X_W), and features of the normal speech (X_S), we have used traditional method of training the GAN as suggested in [16] for our Incpetion-GAN. Inception-GAN consists of one generator (i.e., G_{WS}), and one discriminator (i.e., D_S). The G_{WS} maps input features of the whispered speech to features of the normal speech, and D_S tries to distinguish between the generated speech, and the original normal speech.

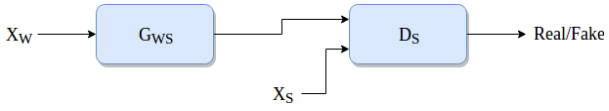

Figure 3: *Schematic representation of proposed Inception-GAN. After [16].*

We have applied adversarial loss function for training the generator since we want to generate the speech such that discriminator gets confused, whether it is real or generated. For training the discriminator, we have used adversarial loss as well since the discriminator should be able to identify real and generated speeches efficiently. Both the loss functions are mathematically represented as:

$$\mathcal{L}_G = -\lambda_1 * [\log(D_S(G_{WS}(X_W)))], \tag{1}$$

$$\mathcal{L}_D = -\lambda_2 * [\log(D_S(X_S)) + \log(1 - D_S(G_{WS}(X_W))], \tag{2}$$

where λ_1 and λ_2 are hyperparameters which decide relative importance of generator, and discriminator loss functions, respectively.

3. Experimental Results

3.1. Experimental Setup

In this paper, we have used whispered TIMIT (wTIMIT) database for the training and testing of our architecture [30]. In particular, we took three male and three female speakers' data for the development of WHSP2SPCH conversion systems. We have in total 1164 parallel utterances corresponding to the

whispered, and the normal speeches are taken for training, and 35 utterances for testing. Each architecture is used to find mapping function between the cepstral features of whispered and normal speech and to find prediction function for the converted features and the F_0 of the normal speech, which is followed by post-processing using a *sinc* interpolation smoothing in the voiced regions of speech.

We have designed CNN-GAN for the baseline architecture. To be fair, we have designed it in such a way that computational complexity of both CNN-GAN, and Inception-GAN is almost similar. The generator and discriminator of CNN-GAN contains four 3×3 convolutions having output number of filters 128, 256, 128, and 1, respectively. Additionally, we have used two fully-connected layers with one dropout of 50% for generator and discriminator, same as generator and discriminator of Inception-GAN. Rectified Linear Unit (ReLU) is also used as an activation function [31]. Both the models are trained for 100 epochs with learning rate of 0.0001 with Adam optimizer [32]. With the window of 25 ms and 5 ms of frame shift, we extracted *40*-dimensional Mel Cepstral Coefficients (MCCs). For analysis and synthesis, we have used AHOCODER [33]. For only training purpose, we have fixed the input size of MCC feature matrix of 1000×40 as suggested in [19]. While testing, number of frames will be variable. For example, we will have $x \times 40$ size of matrix as an input while testing, where x is the number of frames.

3.2. Objective Evaluation

We have applied Mel Cepstral Distortion (MCD), and Root Mean Square Error (RMSE) of $\log(F_0)$-based objective measures to analyze the effectiveness of the WHSP2SPCH conversion systems. The traditional MCD measure is used here which is given by [34]:

$$MCD \text{ [in dB]} = \frac{10}{ln10}\sqrt{2\sum_{i=1}^{40}(m_i^r - m_i^c)^2} \quad , \qquad (3)$$

where m_i^r and m_i^c are the i^{th} MCCs of the reference, and the converted signal. In particular, m_i^r and m_i^c are the i^{th} MCCs of the reference neutral speech, and the converted neutral speech in the case of the WHSP2SPCH conversion system. Since MCD is the distance between the converted and the reference cepstral features, a system that is having lesser MCD is considered as a better system [34].

To measure the RMSE of $\log(F_0)$, the actual reference speech, and the converted speech signals, are time-aligned using the Dynamic Time Warping (DTW) algorithm. These DTW aligned pairs will generate voiced-voiced, voiced-unvoiced, unvoiced-voiced, and unvoiced-unvoiced pairs. Here, we consider only voiced-voiced pairs for computing the RMSE of the $\log(F_0)$ (since F_0 is undefined for the unvoiced frames primarily due to the absence of voicing) [35]. RMSE of the $\log(F_0)$ is given by:

$$RMSE(\log(F_0)) = \sqrt{\sum_{i=1}^{k}[\log(F_{0_i}^r) - \log(F_{0_i}^c)]^2}, \qquad (4)$$

where k is the total number of voiced-voiced pairs after the alignment, ${F_0}^r$ and ${F_0}^c$ are the F_0 of the reference, and the converted speech signals, respectively. Lesser the RMSE of $\log(F_0)$, better the system is.

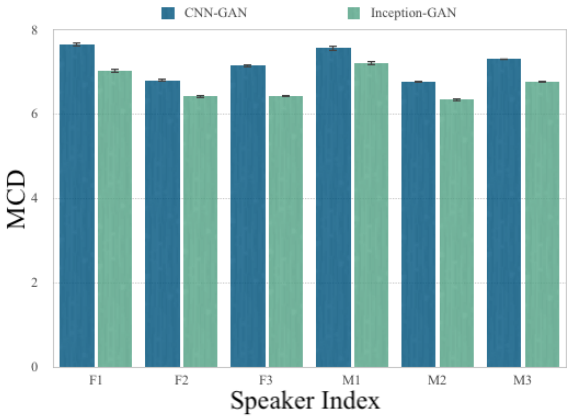

Figure 4: *MCD analysis of the different systems based on parallel WHSP2SPCH task with* 95% *confidence interval. X-axis: Speakers, Y-axis: MCD values.*

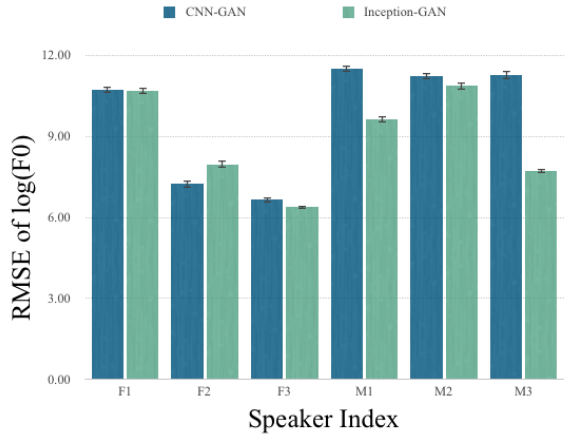

Figure 5: *Analysis of RMSE of* $\log(F_0)$ *for the different systems based on parallel WHSP2SPCH task with* 95% *confidence interval. X-axis: Speakers, Y-axis: RMSE of* $\log(F_0)$ *values.*

We have analyzed the performance of Inception-GAN based on values RMSE $log(F_0)$ and MCD, and compared it with the baseline CNN-GAN. As discussed in Section 2.1, Inception reduces computational complexity which helps increasing the depth and width of the architecture. Moreover, Inception is able to use filter at sparse-level to extract information. These ability of Inception module helps our Inception-GAN to outperform the CNN-GAN. For the MCD analysis, Inception-GAN shows the reduction in MCD compared to the CNN-GAN for all three male, and female speakers, respectively. In addition, for the case of RMSE $log(F_0)$, Inception-GAN performs comparable. All of these results supports all the hypotheses that we discussed in Section 2.

3.3. Convergence

As shown in Figure 6, the loss function output on the validation dataset is increasing for the discriminator, and the generator of the CNN-GAN. In addition, it never converges even at 100 epochs. However, in case of proposed Inception-GAN both

(i.e., generator and discriminator) converges very fast compared to CNN-GAN.

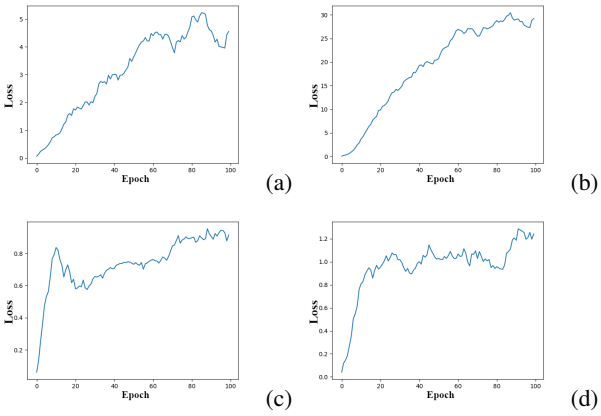

Figure 6: *(a), (b) are the plots of loss functions of discriminator, and generator, respectively, of CNN-GAN. And (c), (d) are the plots of loss functions of discriminator, and generator, respectively, of Inception-GAN.*

3.4. Plots of Global Variance (GV)

We plot the Global Variance (GV) in order to see how much the generated MCC features via CNN-GAN, and Inception-GAN are varying from the target (i.e., GV of original normal speech). Figure 7 and 8 show the GV plots for the male and female speakers, respectively.

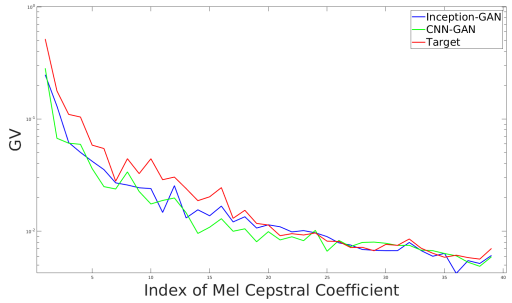

Figure 7: *GV plot of male speaker.*

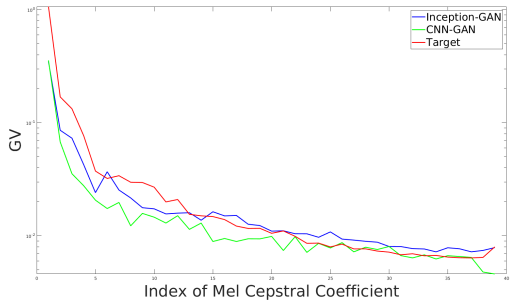

Figure 8: *GV plot of female speaker.*

Mean deviation of CNN-GAN and Inception-GAN based generated speech frm the target speech is 0.3014 and 0.2309, respectively. However, in case of female speakers CNN-GAN and Inception-GAN shows mean deviation of 0.3368 and 0.2280 from target speaker, respectively. We can clearly see that the proposed Inception-GAN is generating the MCC features more like a natural speech.

3.5. Subjective Evaluation

Comparative subjective analysis test, namely, Mean Opinion Score (MOS) has been taken for the subjective evaluations to check the naturalness of the converted speech. Total 16 subjects (7 females and 9 males between 18 to 30 years of age and with no known hearing impairments) took part in the subjective test. Here, we randomly played utterances from both the proposed systems. In the MOS test, subjects were asked to rate the played utterance on the scale of 1-5, where 1 means not at all converted in normal speech, and 5 means completely converted in normal speech. Results of the MOS obtained from the total 128 samples are shown in Figure 9. We can observe that the proposed Inception-GAN is almost 16.67 % (on an average) times more preferred over the CNN-GAN by the subjects.

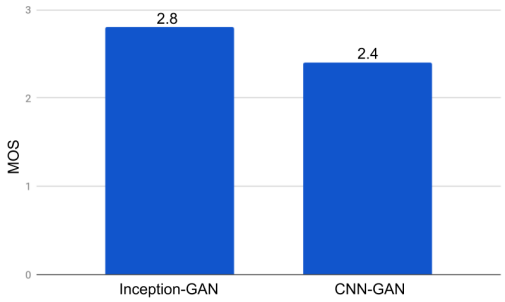

Figure 9: *MOS score analysis for the developed WHSP2SPCH conversion systems.*

4. Summary and Conclusions

In this paper, we analyzed our proposed Inception-GAN for the WHSP2SPCH conversion using parallel data w.r.t. CNN-GAN. Here, we found the limitations of CNN-GAN in terms of computational complexity and time constraints. Moreover, we encountered the problem of overfitting in CNN-GAN. To overcome these limitations, we proposed Inception-GAN which consists of Inception modules in both generator and discriminator. This Inception module has less computational complexity compared to the traditional CNN in GAN. In addition, it also map features at sparse-level which helps us in building more dense architectures. This ability leads to less overfitting on data. In future, we plan to use high-quality vocoders, such as WORLD or Wavenet for further improvement of the voice quality of converted voices. The perceptual difference observed between the estimated and the ground truth indicates the need of exploring the better objective function that can perceptually optimize the network parameters of GAN-based architectures.

5. Acknowledgements

The authors would like to thank the authorities of DA-IICT, Gandhinagar, India and Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India for their kind support to carry out this research work.

6. References

[1] C. Zhang and J. H. L. Hansen, *Advancements in whispered speech detection for interactive/speech systems*. Hemant A. Patil et. al. (Eds), Signal and Acoustic Modelling for Speech and Communication Disorders, De Gruyter, vol. 5, pp. 9–32, 2018.

[2] L. Sulica, "Vocal fold paresis: An evolving clinical concept," *Current Otorhinolaryngology Reports*, vol. 1, no. 3, pp. 158–162, 2013.

[3] A. D. Rubin and R. T. Sataloff, "Vocal fold paresis and paralysis," *Otolaryngologic Clinics of North America*, vol. 40, no. 5, pp. 1109–1131, 2007.

[4] L. Wallis, C. Jackson-Menaldi, W. Holland, and A. Giraldo, "Vocal fold nodule vs. vocal fold polyp: Answer from surgical pathologist and voice pathologist point of view," *Journal of Voice*, vol. 18, no. 1, pp. 125–129, 2004.

[5] J. A. Mattiske, J. M. Oates, and K. M. Greenwood, "Vocal problems among teachers: A review of prevalence, causes, prevention, and treatment," *Journal of Voice*, vol. 12, no. 4, pp. 489–499, 1998.

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

[7] N. Shah, M. Parmar, N. Shah, and H. A. Patil, "Novel mmse discogan for cross-domain whisper-to-speech conversion," in *Machine Learning in Speech and Language Processing (MLSLP) Workshop*, Google Office, Hyderabad, India, 2018, pp. 1–3.

[8] A. Illa, P. K. Ghosh *et al.*, "A comparative study of acoustic-to-articulatory inversion for neutral and whispered speech," in *ICASSP*, New Orleans, USA, 2017, pp. 5075–5079.

[9] N. J. Shah and H. A. Patil, *Non-audible murmur to audible speech conversion*. Voice Technologies for Reconstruction and Enhancement, Hemant A. Patil and A. Neustein (Eds), De Gruyter, vol.6, 2019.

[10] G. Srinivasan, A. Illa, and P. K. Ghosh, "A study on robustness of articulatory features for automatic speech recognition of neutral and whispered speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5936–5940.

[11] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education India, 1^{st} (Eds.), 2006.

[12] H. Konno, M. Kudo, H. Imai, and M. Sugimoto, "Whisper-to-normal speech conversion using pitch estimated from spectrum," *Speech Communication*, vol. 83, pp. 10–20, 2016.

[13] W. Meyer-Eppler, "Realization of prosodic features in whispered speech," *The J. of the Acoust. Soc. of Amer. (JASA)*, vol. 29, no. 1, pp. 104–106, 1957.

[14] T. Itoh, K. Takeda, and F. Itakura, "Acoustic analysis and recognition of whispered speech," in *ASRU*, Madonna di Campiglio, Italy, 2001, pp. 429–432.

[15] M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad, "Fundamental frequency generation for whisper-to-audible speech conversion," in *ICASSP*, Florence, Italy, 2014, pp. 2579–2583.

[16] 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.

[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. 3642–3646.

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

[19] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *ICASSP*, Calgary, Alberta, Canada, 2018, pp. 5039–5043.

[20] 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.

[21] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 1, pp. 84–96, 2018.

[22] N. Shah, R. Sreeraj, N. Shah, and H. A. Patil, "Novel inter mixture weighted GMM posteriorgram for dnn and gan-based voice conversion," to appear in *Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*, Honolulu, Hawaii, USA, 2018.

[23] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversariel network," in *ICASSP*, Calgary, Alberta, Canada, 2018, pp. 5279–5283.

[24] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018, pp. 2100–2104.

[25] S. Seshadri, L. Juvela, J. Yamagishi, O. Rasanen, and P. Alku, "Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion," in *ICASSP*, Brighton, UK, 2019.

[26] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.

[27] N. J. Shah, M. Parmar, N. Shah, and H. A. Patil, "Novel mmse discogan for cross-domain whisper-to-speech conversion," in *Machine Learning in Speech and Language Processing (MLSLP) Workshop*, 2018.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[30] B. P. Lim, *Computational differences between whispered and non-whispered speech*. Ph.D. Thesis, University of Illinois at Urbana-Champaign, USA, 2011.

[31] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean *et al.*, "On rectified linear units for speech processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, British Columbia, Canada, 2013, pp. 3517–3521.

[32] D. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *International Confernece on Learning Representation (ICLR)*, San Diego, USA, 2015, pp. 1–15.

[33] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Improved HNM-based vocoder for statistical synthesizers," in *INTERSPEECH*, Florence, Italy, 2011, pp. 1809–1812.

[34] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. on Audio, Speech and Lang. Process. (TASLP)*, vol. 15, no. 8, pp. 2222–2235, 2007.

[35] Z.-Z. Wu, T. Kinnunen, E. S. Chng, and H. Li, "Text-independent F_0 transformation with non-parallel data for voice conversion," in *INTERSPEECH*, Makuhari, Chiba, Japan, 2010, pp. 1732–1735.