

CinC-GAN for Effective F_0 prediction for Whisper-to-Normal Speech Conversion

Maitreya Patel, Mirali Purohit, Jui Shah, and Hemant A. Patil

Speech Research Lab, DA-IICT, Gandhinagar-382007, India.

E-mail: {maitreya_patel, purohit_mirali, jui_shah, hemant_patil}@daiict.ac.in

Abstract—Recently, Generative Adversarial Networks (GAN)-based methods have shown remarkable performance for the Voice Conversion and WHiSPer-to-normal SPeeCH (WHSP2SPCH) conversion. One of the key challenges in WHSP2SPCH conversion is the prediction of fundamental frequency (F_0). Recently, authors have proposed state-of-the-art method Cycle-Consistent Generative Adversarial Networks (CycleGAN) for WHSP2SPCH conversion. The CycleGAN-based method uses two different models, one for Mel Cepstral Coefficients (MCC) mapping, and another for F_0 prediction, where F_0 is highly dependent on the pre-trained model of MCC mapping. This leads to additional non-linear noise in predicted F_0 . To suppress this noise, we propose Cycle-in-Cycle GAN (i.e., CinC-GAN). It is specially designed to increase the effectiveness in F_0 prediction without losing the accuracy of MCC mapping. We evaluated the proposed method on a non-parallel setting and analyzed on speaker-specific, and gender-specific tasks. The objective and subjective tests show that CinC-GAN significantly outperforms the CycleGAN. In addition, we analyze the CycleGAN and CinC-GAN for unseen speakers and the results show the clear superiority of CinC-GAN.

Index Terms—Whisper-to-Normal Speech, Non-parallel, F_0 prediction, CycleGAN, CinC-GAN.

I. INTRODUCTION

Whisper and normal speech are different way of communication. People generally use normal mode of speech in regular life, however in some cases, people need to keep their conversation private such as, during phone calls in public places, in meeting, library, hospital etc, where people adopt to use whisper mode conversation [1]. Whisper and normal speech are cross-domain entities, as it differs in terms of speech production and perception [1]–[3]. Given a speech, whether it is normal or not is depend on arrangements of larynx, and particularly on glottis [4]–[7]. Sometimes because of accident or disease, people are not able to produce normal speech, because the parts which take part in speech production get affected. Also losing the normal way of speaking will significantly affect the person’s life. When people speak in normal style, vocal folds vibrates with some specific fundamental frequency (i.e., F_0) while this is not the case in whisper speech [1], [8]. In addition, current speech processing systems do not perform efficiently on any kind of speech except on normal speech. Therefore, WHSP2SPCH conversion task is necessary.

One of the challenging problem in WHSP2SPCH conversion is F_0 prediction. However, F_0 is encapsulated in an intricate way in the whispered speech. The presence and absence of F_0 is the key difference between normal vs. whis-

pered speech [9]–[11]. At the acoustic-level, there is difference between voiced and unvoiced speech, and statistical voice conversion (VC)-based methods are able to do such conversion [12]. Attempts have been made in the literature for VC, such as GMM, Conditional Variational AutoEncoders (CVAE), CycleGAN-VC, etc. [13]–[18]. For WHSP2SPCH conversion attempts have been made in the literature using parallel data only. Such as LSTM, MSpeC-Net, DiscoGAN, CycleGAN, etc. are proposed in the literature [2], [9], [12], [19]–[24]. Moreover, CycleGAN has shown state-of-the-art result for WHSP2SPCH conversion including F_0 prediction on parallel data, which relies on the availability of particular speaker’s whisper, and normal speech [25]. However, this is not feasible and it is impractical too. Moreover, parallel data requires time-alignment as pre-processing. In addition, traditional method uses 2-step sequential method for WHSP2SPCH conversion [25], [26]. For CycleGAN based conversion, in first step, one CycleGAN is trained for cepstral feature mapping of whisper to normal speech, and in second step, another CycleGAN is trained for F_0 prediction, which heavily relies on previously trained CycleGAN [25]. Because of the imperfect cepstral feature mapping, noise is introduced in the output. Due to the non-linear DNN layers, it is non-linear noise. Therefore, significant non-linear noise is added in F_0 prediction.

Although CycleGAN gives the state-of-the-art result, there is still a gap between the original and converted normal speech in terms of naturalness [25]. To reduce this gap and overcome above limitations, we propose CinC-GAN for non-parallel WHSP2SPCH conversion task, including F_0 prediction in non-parallel mode. CinC-GAN is designed specifically for effective F_0 prediction, which is important factor for naturalness. Here, CinC-GAN uses joint training methodology, where acoustic mapping, and F_0 prediction is done simultaneously. The objective result shows that CinC-GAN is able to suppress the non-linear noise in F_0 prediction. Therefore, F_0 -RMSE is decreased by 29.8% and 82.2% compared to the baseline for speaker and gender-specific tasks, respectively. Subjective evaluation shows that CinC-GAN helps to bring the converted normal speech more closer to the original normal speech compared to the baseline (CycleGAN). In objective and subjective evaluations, gender-specific task contains analysis on seen and unseen speakers. In addition, CinC-GAN maintains the naturalness for gender-specific task (for seen and unseen speakers), whereas CycleGAN degrades its result and produces whisper speech.

II. CONVENTIONAL CYCLE-GAN

Let $x \in R^N$ and $y \in R^N$ be the cepstral features of whisper (X) and normal (Y) speech, respectively, where N is the dimension of a feature vector. In CycleGAN, two generators are used, $G_{X \rightarrow Y}$ and $G_{Y \rightarrow X}$, where $G_{X \rightarrow Y}$ maps the cepstral features of X to Y , whereas mapping $G_{Y \rightarrow X}$ does the opposite (i.e., Y to X). In addition, we have two discriminators D_X and D_Y , whose role is to predict whether its input is from the distribution X and Y or not, respectively.

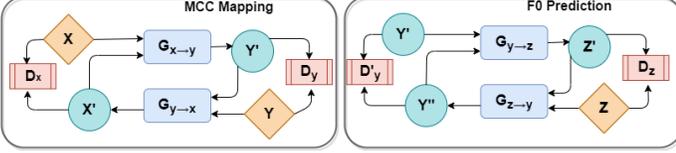


Fig. 1: Conventional CycleGAN. After [27].

In CycleGAN, there are three types of losses, cycle-consistent loss, adversarial loss, and identity loss, as described below.

Adversarial loss: To make converted normal speech indistinguishable from the original, we use adversarial loss. Here, we use least square error loss instead of traditional binary cross-entropy loss, which is defined as:

$$\mathcal{L}_{adv}(G_{X \rightarrow Y}, D_Y) = \mathbb{E}_{y \sim P_Y(y)} [(D_Y(y) - 1)^2] + \mathbb{E}_{x \sim P_X(x)} [(D_Y(G_{X \rightarrow Y}(x)))^2]. \quad (1)$$

Cycle-consistent loss: The main idea behind this loss is to map the distribution between the original and reconstructed data. In addition, this loss tries to preserve contextual information across different speech. This loss allows us to do non-parallel WHSP2SPCH conversion. The loss is defined as:

$$\begin{aligned} \mathcal{L}_{cyc}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) &= \mathbb{E}_{x \sim P_X(x)} [\|G_{Y \rightarrow X}(G_{X \rightarrow Y}(x)) - x\|_1] \\ &+ \mathbb{E}_{y \sim P_Y(y)} [\|G_{X \rightarrow Y}(G_{Y \rightarrow X}(y)) - y\|_1]. \end{aligned} \quad (2)$$

Identity-mapping loss: To encourage preservation of input linguistic content (as suggested in [27]), identity loss is used:

$$\mathcal{L}_{id}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) = \mathbb{E}_{x \sim P_X(x)} [\|G_{Y \rightarrow X}(x) - x\|_1] + \mathbb{E}_{y \sim P_Y(y)} [\|G_{X \rightarrow Y}(y) - y\|_1]. \quad (3)$$

The total loss function is defined as:

$$\begin{aligned} \mathcal{L}_{full} &= \mathcal{L}_{adv}(G_{X \rightarrow Y}, D_Y) + \mathcal{L}_{adv}(G_{Y \rightarrow X}, D_X) \\ &+ \lambda_{cyc} \mathcal{L}_{cyc}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) + \lambda_{id} \mathcal{L}_{id}(G_{X \rightarrow Y}, G_{Y \rightarrow X}). \end{aligned} \quad (4)$$

Where the values of λ_{cyc} and λ_{id} are 10 and 5, respectively. Now, for F_0 prediction, we train another CycleGAN architecture, where $y' \in R^N$ is the cepstral features of converted normal, which is extracted from previously trained CycleGAN for MCC mapping, and $z \in R^1$ is the F_0 of original normal speech.

III. PROPOSED CINC-GAN

Problem formulation: The conventional formulation for WHSP2SPCH conversion is $y' = f(x) + n$ for cepstral feature mapping, where x is whisper speech features, f is the mapping function, and n is the additive noise. Now, for F_0 prediction, we formulate the problem as $z = g(y') + n'$, which implies that $z = g(f(x) + n) + n'$, where g is the mapping function, and n' another additive noise.

Given this problem, we observed that due to the use of two differently trained mapping functions, for F_0 prediction, significant non-linear noise is being added. Hence, for effective F_0 prediction and to suppress this noise, we need some sophisticated mapping function, which can be trained simultaneously, and somehow it can also directly rely on input instead of only $f(x)$.

Proposed solution: In this paper, we propose a different training method, namely, Cycle-in-Cycle GAN (CinC-GAN), which is an advanced version of CycleGAN, for WHSP2SPCH conversion. In CycleGAN, we use one model for acoustic feature mapping, and second for F_0 prediction, where both of them are separately trained (i.e., sequential training). However, in CinC-GAN, we use inner cycle for acoustic feature mapping, and outer cycle for F_0 prediction, where outer cycle relies on cepstral features of converted normal speech, and input whisper speech as well (i.e., joint training). This way, we are able to achieve our goal, and suppress the effect of extra noise.

In summary, we propose a Cycle-in-Cycle GAN as shown in Fig. 2. In this approach, we adopt two coupled CycleGANs to learn the mapping for X to Y and Y to Z , respectively. In addition, non-parallel dataset $x \in X$, $y \in Y$, and $z \in Z$ is used for training, where X and Y are set of cepstral features of whisper and normal speech, respectively, and Z is set of F_0 extracted from the normal speech. Detailed description on feature extraction is given in Section IV.

A. Acoustic feature mapping

The inner cycle in Fig. (2) maps cepstral features of whisper (X) to normal speech (Y). We use two generators, $G_{X \rightarrow Y}$ and $G_{Y \rightarrow X}$, where $G_{X \rightarrow Y}$ maps x to Y and $G_{Y \rightarrow X}$ maps y to X . The discriminators D_X and D_Y confirms whether generated distribution is from X and Y or not, respectively. Here, we use adversarial loss, cycle-consistency loss, and identity loss. Adversarial loss is defined as:

$$\mathcal{L}_{adv_1}(G_{X \rightarrow Y}, D_Y) = \mathbb{E}_{y \sim P_Y(y)} [(D_Y(y) - 1)^2] + \mathbb{E}_{x \sim P_X(x)} [(D_Y(G_{X \rightarrow Y}(x)))^2]. \quad (5)$$

To map the two different distributions (i.e., normal and whisper speech), we add generator $G_{Y \rightarrow X}$ to map normal-to-whisper speech features. In addition, we use discriminator, D_X to distinguish between real and generated whisper speech. Therefore, we also use single cycle-consistency loss: i.e.,

$$\mathcal{L}_{cyc_1}(G_{X \rightarrow Y}) = \mathbb{E}_{x \sim P_X(x)} [\|G_{Y \rightarrow X}(G_{X \rightarrow Y}(x)) - x\|_1]. \quad (6)$$

In addition, we use identity loss to preserve the linguistic content, i.e.,

$$\mathcal{L}_{id_1}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) = \mathbb{E}_{x \sim P_X(x)} [\|G_{Y \rightarrow X}(x) - x\|_1] + \mathbb{E}_{y \sim P_Y(y)} [\|G_{X \rightarrow Y}(y) - y\|_1]. \quad (7)$$

B. F_0 Prediction

After mapping the cepstral features of whisper-to-normal speech, we focus on F_0 prediction task. Previous methods tries to predict F_0 from the cepstral features of converted normal speech using CycleGAN, which is trained separately (i.e., sequential training). However, in this paper, we propose to predict F_0 from the cepstral features of converted normal speech simultaneously via joint training.

We use the generator $G_{Y \rightarrow Z}$ to predict F_0 from the converted normal speech ($G_{X \rightarrow Y}(x)$) and $G_{Z \rightarrow X}$ is used to map the predicted F_0 to whisper speech instead of normal speech. This way, we are able to remove the non-linear noise by including the effect of original whisper speech and joint training methodology. In addition, we use discriminator D_Z to make generated F_0 just like original F_0 . However, to add the effect of whisper speech, we used fourth generator to generate whisper speech features from the predicted F_0 instead to converted normal speech features. Here, we adapt only two losses, adversarial loss, and cycle-consistency loss, i.e.,

$$\mathcal{L}_{adv_2}(G_{Y \rightarrow Z}, D_Z) = \mathbb{E}_{z \sim P_Z(z)} [(D_Z(z) - 1)^2] + \mathbb{E}_{y \sim P_Y(y)} [(D_Z(G_{Y \rightarrow Z}(y)))^2]. \quad (8)$$

$$\mathcal{L}_{cyc_2}(G_{Y \rightarrow Z}) = \mathbb{E}_{y \sim P_Y(y)} [\|G_{Z \rightarrow Y}(G_{Y \rightarrow Z}(y)) - y\|_1]. \quad (9)$$

Moreover, we add combine loss through a third discriminator D_X . This discriminator confirms the output of two generators ($G_{Y \rightarrow X}$, $G_{Z \rightarrow X}$) is from original distribution of X or not. This way both (inner and outer) cycles stay connected with common measure of reconstruction.

$$\begin{aligned} \mathcal{L}_{adv_3}(G_{Y \rightarrow X}, G_{Z \rightarrow X}) &= \mathbb{E}_{x \sim P_X(x)} [(D_X(x) - 1)^2] \\ &+ \mathbb{E}_{y \sim P_Y(y)} [(D_X(G_{Y \rightarrow X}(G_{X \rightarrow Y}(x))))^2] \\ &+ \mathbb{E}_{z \sim P_Z(z)} [(D_X(G_{Z \rightarrow X}(G_{Y \rightarrow Z}(G_{X \rightarrow Y}(x))))^2]. \end{aligned} \quad (10)$$

C. Overall Objective of the Proposed Method

In summary, we train both the cycles simultaneously. And we optimize all the generators, and discriminators according to the following rules:

$$\mathcal{L}_{full} = \mathcal{L}_{adv_1} + \lambda_1 * \mathcal{L}_{cyc_1} + \lambda_2 * \mathcal{L}_{id_1} + \lambda_3 * \mathcal{L}_{adv_2} + \lambda_4 * \mathcal{L}_{cyc_2} + \lambda_5 * \mathcal{L}_{adv_3}, \quad (11)$$

where λ_1 , λ_2 , λ_3 , λ_4 , and λ_5 are the hyperparameters associated with different loss functions. These parameters defines relative importance of each losses w.r.t. the other

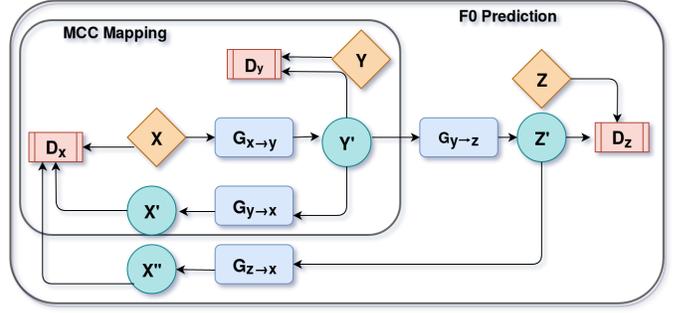


Fig. 2: Proposed Cycle-in-Cycle GAN. After [28].

losses. Here, $\lambda_1 = 10$, $\lambda_2 = 5$, $\lambda_3 = 10$, $\lambda_4 = 1$, and $\lambda_5 = 1$ are used *empirically* in all of our experiments (because this choice of hyperparameters shows stable and accurate training). And these hyperparameter values work for any conversion pairs.

IV. EXPERIMENTAL RESULTS

A. Dataset and Feature Extraction

In WHSP2SPCH conversion, we have used Whispered TIMIT (wTIMIT) database [29]. In both the approaches, i.e., speaker-specific and gender-specific, we have done non-parallel training. We have done speaker-specific task on four different speakers, specifically two female and two male speakers. Particularly, for each speaker, 34 minutes of training data and 2.32 minutes of testing data was used. In each gender-specific task, we have used four speakers, and particularly, in each training, 136 minutes of training data was used. In gender-specific task, we test it on four seen and two unseen speakers, and test data for each speaker is 13.92 minutes. We extract the F_0 and MCC (Mel Cepstral Coefficient) features from whisper and normal speech using AHOCODER [30]. In feature extraction, we have used 25 ms window size, and 5 ms frame shift [30].

B. Architecture Details

Generators $G_{X \rightarrow Y}$, $G_{Y \rightarrow X}$ and $G_{Y \rightarrow Z}$ follow the same configuration, for both the architectures. In $G_{X \rightarrow Y}$ and $G_{Y \rightarrow X}$, contain 40, 512, and 40 neurons in input layer, hidden layers, and output layer, respectively. Generator $G_{Y \rightarrow Z}$ contains 40, 512, and 1 neurons in input layer, hidden layers and output layer, respectively. $G_{Z \rightarrow Y}$ has the 1, 512, and 40 neurons in input layer, hidden layers, and output layer, respectively. All layers are followed by Rectified Linear Unit (ReLU) activation function. All discriminators follow the same configuration for both the architecture. D_X , D_Y , and D'_Y have the 40, 512, and 1 neurons in the input layer, hidden layers, and output layer, respectively. D_Z has the 1, 512, and 1 neurons in the input layer, hidden layers, and output layer, respectively. In all discriminators, input layer and all hidden layers are followed by *ReLU* activation function and output layer followed by *sigmoid* activation function. Both the architectures are trained for 100 epochs,

TABLE I: MCD analysis of the different WHSP2SPCH systems for speaker-specific task. Here, % in the bracket indicates the relative reduction in the MCD w.r.t the baseline

| Method \ Speaker | F1 (US_102) | M1 (US_103) | F2 (US_104) | M2 (US_106) |
|---------------------|-------------|---------------|--------------|-------------|
| CycleGAN (Baseline) | 6.76 | 6.36 | 6.1 | 5.97 |
| CinC-GAN | 6.73 (0.4%) | 6.42 (-0.94%) | 6.11 (-0.1%) | 5.86 (2%) |

TABLE II: MCD analysis of the different WHSP2SPCH systems for gender-specific task. Here, % in the bracket indicates the relative reduction in the MCD w.r.t the baseline

| Method \ Speaker | F-Seen | M-Seen | F-Unseen | M-Unseen |
|---------------------|--------------|---------------|--------------|-------------|
| CycleGAN (Baseline) | 6.69 | 6.28 | 6.77 | 6.83 |
| CinC-GAN | 6.66 (0.45%) | 6.29 (-0.16%) | 6.92 (-2.2%) | 6.9 (-1.0%) |

and learning rate was set to 0.0001. Source code is provided at <https://github.com/Maitreyapatel/speech-conversion-between-different-modalities>.

C. 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 [31]. 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. Lesser the RMSE of $\log(F_0)$, better the system is.

The effectiveness of CinC-GAN can be clearly seen for the WHSP2SPCH conversion system in objective results. Analysis of both the architectures is done using 2 different approaches 1) speaker-specific in which model is trained and tested only on single speaker and 2) gender-specific in which model is trained for specific number of speakers and tested on seen as well as out of the box speaker (unseen speaker). As shown in Table I, it can be observed that CinC-GAN performs comparatively to CycleGAN in terms of MCD. However, CinC-GAN outperforms CycleGAN in terms of RMSE $\log(F_0)$ for all the speakers (as shown in Table III). CinC-GAN gets on an average 29.8% relative reduction in case of speaker-specific, compared to the CycleGAN in F_0 -RMSE. Moreover, Table V shows the Kullback-Leibler Divergence (KLD) and Jensen-Shannon Divergence (JSD) between predicted F_0 and original F_0 for speaker-specific task. Here, we can observe that CinC-GAN outperforms CycleGAN. Therefore, this analysis further strengthens our results.

TABLE V: Results of KL-JSD for Speaker-specific task.

| Speaker | Method | | CycleGAN | |
|---------|----------|----------|----------|------|
| | CinC-GAN | CycleGAN | KL | JSD |
| US_102 | 5.11 | 4.95 | 29.23 | 4.76 |
| US_103 | 5.27 | 5.83 | 0.03 | 7.22 |
| US_104 | 7.29 | 4.46 | 20.94 | 7.74 |
| US_106 | 2.37 | 3.59 | 4.85 | 1.27 |
| Average | 7.51 | 4.71 | 13.76 | 5.25 |

TABLE III: RMSE-based objective analysis of $\log(F_0)$ for speaker-specific task. Here, % in the bracket indicates a relative reduction in the RMSE w.r.t the baseline

| Method \ Speaker | F1 (US_102) | M1 (US_103) | F2 (US_104) | M2 (US_106) |
|---------------------|--------------|-------------|--------------|--------------|
| CycleGAN (Baseline) | 7.19 | 5.7 | 3.88 | 6.49 |
| CinC-GAN | 5.65 (21.4%) | 4.6 (19.3%) | 2.77 (28.5%) | 3.25 (49.9%) |

TABLE IV: RMSE-based objective analysis of $\log(F_0)$ for gender-specific task. Here, % in the bracket indicates a relative reduction in the RMSE w.r.t the baseline

| Method \ Speaker | F-Seen | M-Seen | F-Unseen | M-Unseen |
|---------------------|--------------|--------------|--------------|-------------|
| CycleGAN (Baseline) | 18.2 | 38.9 | 25.6 | 28.3 |
| CinC-GAN | 6.81 (62.3%) | 3.16 (91.9%) | 3.14 (87.7%) | 3.8 (86.6%) |

We further extend our experiment, and perform objective evaluation for gender-specific task. For this, we trained two CinC-GAN, first on 4 female speakers, and second on 4 male speakers. We tested both of them on seen speaker and unseen utterances, and unseen speaker, as well. As shown in Table II, in terms of MCD, CycleGAN and CinC-GAN performs similarly. However, in terms of F_0 -RMSE CinC-GAN outperforms CycleGAN by on an average 82.1%, as shown in Table IV. We observed that the CycleGAN is not able to predict F_0 effectively on combined dataset, whereas CinC-GAN works quite efficiently in every scenarios even on unseen speaker and unseen utterances.

D. Subjective Evaluation

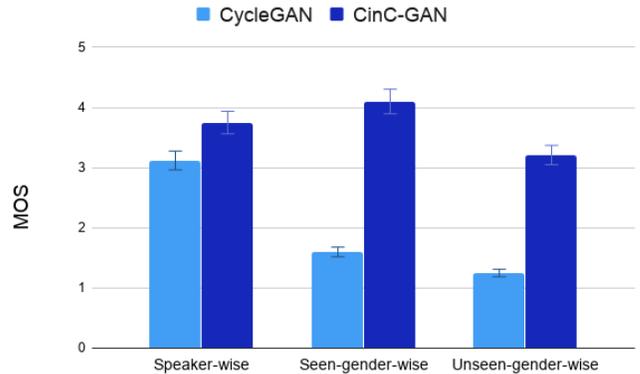


Fig. 3: MOS score analysis for speaker-specific and gender-specific task (i.e., seen-unseen) with 95% confidence interval.

For subjective test analysis, Mean Opinion Score (MOS) has been taken to measure the naturalness of the converted speech. Total 28 subjects (7 females and 21 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 systems. In the MOS test, subjects were asked to rate the played utterances on the scale of 1-5, where 1 indicates completely whisper speech, and 5 means completely converted in normal speech. We can observe that the CinC-GAN has

almost 20.2% more naturalness in case of speaker-specific task. From Fig. 3, we can observe that CinC-GAN significantly outperforms CycleGAN for seen and unseen (out-of-the-box) speakers, respectively, on gender-specific task. In addition, in this case, CycleGAN fails measurably and produces whisper speech even for seen and unseen speakers, which can be observed in MOS plot shown in Fig. 3. However, CinC-GAN maintains its performance for seen and unseen speakers. CinC-GAN is able to score $MOS \geq 3$ for gender-specific task for unseen speaker as well. Therefore, CinC-GAN leads to the possibility of few-shot learning for WHSP2SPCH for the first time in literature.

V. SUMMARY AND CONCLUSION

In this paper, we proposed the CinC-GAN to increase the effectiveness of F_0 prediction without affecting accuracy of MCC mapping. Baseline (i.e., CycleGAN) uses sequential training, which adds non-linear noise in F_0 prediction. However, CinC-GAN adopts joint training methodology to decrease this noise. Objective and subjective results show superiority of CinC-GAN over the baseline. In addition, CycleGAN fails in WHSP2SPCH conversion for gender-specific task. However, CinC-GAN maintains its result even for out-of-the-box speaker. This shows the potential of CinC-GAN for few-shot WHSP2SPCH conversion. In future, we plan to extend our study on zero-shot and one-shot WHSP2SPCH conversion.

REFERENCES

- [1] Chi Zhang and John 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] Nirmesh J. Shah, Mihir Parmar, Neil Shah, and Hemant A. Patil, “Novel MMSE DiscoGAN for cross-domain whisper-to-speech conversion,” in *Machine Learning in Speech and Language Processing (MLSPL) Workshop*, Google Office, Hyderabad, India, 2018, pp. 1–3.
- [3] Aravind Illa, Prasanta Kumar Ghosh, et al., “A comparative study of acoustic-to-articulatory inversion for neutral and whispered speech,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 5075–5079.
- [4] Lesly Wallis, Cristina Jackson-Menaldi, Wayne Holland, and Alvaro 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] Jacqueline A Mattiske, Jennifer M Oates, and Kenneth 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] Lucian Sulica, “Vocal fold paresis: An evolving clinical concept,” *Current Otorhinolaryngology Reports*, vol. 1, no. 3, pp. 158–162, 2013.
- [7] Adam D Rubin and Robert T Sataloff, “Vocal fold paresis and paralysis,” *Otolaryngologic Clinics of North America*, vol. 40, no. 5, pp. 1109–1131, 2007.
- [8] Thomas F Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson Education India, 1st (Eds.), 2006.
- [9] Hideaki Konno, Mineichi Kudo, Hideyuki Imai, and Masanori Sugimoto, “Whisper to normal speech conversion using pitch estimated from spectrum,” *Speech Communication*, vol. 83, pp. 10–20, 2016.
- [10] Werner 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.
- [11] Taisuke Itoh, Kazuya Takeda, and Fumitada Itakura, “Acoustic analysis and recognition of whispered speech,” in *Automatic Speech Recognition and Understanding (ASRU)*, Madonna di Campiglio, Italy, 2001, pp. 429–432.
- [12] Tomoki Toda, Mikihiro Nakagiri, and Kiyohiro Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.
- [13] Yannis Stylianou, Olivier Cappé, and Eric Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.
- [14] Diederik P Kingma and Max Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013, {Last Accessed: May 01, 2014}.
- [15] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai, “Voice conversion using deep neural networks with layer-wise generative training,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1859–1872, 2014.
- [16] Takuhiro Kaneko and Hirokazu Kameoka, “Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks,” in *European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018, pp. 2100–2104.
- [17] Maitreya Patel, Mihir Parmar, Savan Doshi, Nirmesh J. Shah, and Hemant A. Patil, “Novel adaptive generative adversarial network for voice conversion,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Lanzhou, China, 2019, pp. 1273–1281.
- [18] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, “CYCLEGAN-VC2: Improved cyclegan-based non-parallel voice conversion,” in *ICASSP*, Brighton, UK, 2019.
- [19] Harshit Malaviya, Jui Shah, Maitreya Patel, Jalansh Munshi, and Hemant A Patil, “Mspec-net: Multi-domain speech conversion network,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7764–7768.
- [20] G Nisha Meenakshi and Prasanta Kumar Ghosh, “Whispered speech-to-neutral speech conversion using bidirectional LSTMs,” in *INTER-SPEECH*, Hyderabad, India, 2018, pp. 491–495.
- [21] Matthias Janke, Michael Wand, Till Heistermann, Tanja Schultz, and K Prahallad, “Fundamental frequency generation for whisper-to-audible speech conversion,” in *ICASSP*, Florence, Italy, 2014, pp. 2579–2583.
- [22] Ian Vince McLoughlin, Jingjie Li, and Yan Song, “Reconstruction of continuous voiced speech from whispers,” in *INTERSPEECH*, Lyon, France, 2013, pp. 1022–1026.
- [23] Ian V Mcloughlin et al., “Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation,” *ACM Transactions on Accessible Computing (TACCESS)*, vol. 6, no. 4, pp. 12, 2015.
- [24] Viet-Anh Tran, Gérard Bailly, Hélène Løevenbruck, and Tomoki Toda, “Multimodal HMM-based NAM-to-speech conversion,” in *INTER-SPEECH*, Brighton, United Kingdom (UK), 2009, pp. 656–659.
- [25] Mihir Parmar, Savan Doshi, Nirmesh J. Shah, Maitreya Patel, and Hemant A. Patil, “Effectiveness of cross-domain architectures for whisper-to-normal speech conversion,” in *27th European Signal Processing Conference (EUSIPCO)*, Coruña, Spain, 2019.
- [26] Maitreya Patel, Mihir Parmar, Savan Doshi, Nirmesh Shah, and Hemant Patil, “Novel Inception-GAN for Whispered-to-Normal Speech Conversion,” in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 87–92.
- [27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *ICCV*, Venice, Italy, 2017, pp. 1–8.
- [28] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin, “Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 701–710.
- [29] Boon Pang Lim, *Computational differences between Whispered and Non-whispered Speech.*, Ph.D. Thesis, University of Illinois at Urbana-Champaign, USA, 2011.
- [30] D. Erro, I. Sainz, E. Navas, and I. Hernáez, “Improved HNM-based vocoder for statistical synthesizers,” in *INTERSPEECH*, Florence, Italy, 2011, pp. 1809–1812.
- [31] 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.