Effectiveness of Cross-Domain Architectures for Whisper-to-Normal Speech Conversion

Mihir Parmar, Savan Doshi, Nirmesh J. Shah, Maitreya Patel and Hemant A. Patil
Speech Research Lab, DA-IICT, Gandhinagar-382007, India.
E-mail: {mihir_parmar, savan_doshi, nirmesh88_shah, maitreya_patel, hemant_patil}@daiict.ac.in

***Abstract*—Though whisper is a typical way of natural speech communication, it is different from normal speech w.r.t. to speech production and perception perspective. Recently, authors have proposed Generative Adversarial Network (GAN)-based architecture (namely, DiscoGAN) to discover such cross-domain relationships for whisper-to-normal speech (WHSP2SPCH) conversion. In this paper, we extend this study with detailed theory and analysis. In addition, Cycle-consistent Adversarial Network (CycleGAN) is also proposed for the cross-domain WHSP2SPCH conversion. We observe that the proposed systems yield objective results that are comparable to the baseline, and are superior in terms of fundamental frequency (i.e., F_0) prediction. Moreover, we observe that the proposed cross-domain architectures have been preferred 55.75% (on average) times more compared to the traditional GAN in the subjective evaluations. This reveals that the proposed method yields a more natural-sounding normal speech converted from whispered speech.**

***Index Terms*—Whisper, Normal Speech, Cross-domain, GAN, DiscoGAN, CycleGAN.**

I. Introduction

In recent decades, speech technologies have made remarkable progress. However, many barriers still exist in whispered speech applications [1]. Interesting applications of the whispered speech communications are, private conversation in public using cell phone, conversation in quiet environments like a library, a hospital, a meeting room, etc. [1]. Furthermore, the patients that are suffering from the vocal fold paralysis [2], [3], vocal nodule [4], [5], etc. may not be able to produce normal speech due to the partial or complete absence of vocal fold vibrations (i.e., voicing). Losing the natural way of producing the speech will affect one's life extremely, since speech is the most natural and powerful form of communication among humans. Hence, the aim of the present work is to convert whispered speech into normal speech using Machine Learning (ML)-based approaches in order to improve the quality of communication. Attempts have been made in the past to predict fundamental frequency (i.e., F_0) in the WHSP2SPCH conversion [6]–[16]. Though F_0 is absent in speech, it has been observed that the sensation of pitch exists in the whispered speech (which is encapsulated in an intricate way [6], [17]–[19]). Hence, predicting F_0 from the whispered speech is one of the most challenging task.

Though whisper is a usual mode of speech communication, both whispered and normal speeches are different w.r.t the speech production-perception perspective [1], [16], [20]. During the normal speech production, airflow from the lungs is regulated by a periodic vibration of the vocal folds and as a result, voiced sounds are produced. However, during the whispered speech production, vocal folds do not vibrate (i.e., the glottis is opened) which causes exhaled air to pass through the glottal constrictions [1], [21]. This results in the noisy source excitation for the vocal tract system [22]. Hence, the efforts that are put on vocal folds are different for both the types of speeches [1]. In addition, the noise excitation in the whispered speech is *normally* distributed across the lower portion of the vocal tract, which results in *20* dB reduction of the power than its normal speech counterpart [23]. Furthermore, the whispered speech is completely aperiodic or unvoiced in nature due to the lack of any periodic segments [1]. It has also been observed that a change in the overall spectral slope, formant locations (i.e., a shifting of the boundaries of vowel regions in the F1-F2 frequency space), and a change in both energy and duration characteristics in the whispered speech compared to its normal speech counterpart [1], [20], [24]. These temporal and spectral differences significantly reduce the intelligibility of the whispered speech [1], [25]. Hence, the conversion of whispered speech to normal speech is formidable and exigent task [26].

Recently, GAN-based architectures have attracted significant attention for various speech technology problems, such as Speech Enhancement (SE), Voice Conversion (VC), etc. [16], [27]–[31]. It is very natural for humans to acknowledge cross-domain relationships so easily due to their efficient perception mechanism. However, it is difficult for machines to achieve the same ability [32]. The problem is to find a mapping between two domains. The ability of GANs in modeling a latent representation (due to its ability to learn probability density function (*pdf*)) [33], has shown a significant improvement in various Voice Conversion (VC) applications [27]–[31], [34]. As suggested in [16], [32], [35], CycleGAN and DiscoGAN can learn cross-domain relationships without a pair-labeling dataset. Here, we propose the CycleGAN and Mean Square Error (MSE) regularized DiscoGAN (i.e., MMSE DiscoGAN) architectures for the cross-domain WHSP2SPCH conversion task with significant modifications in the loss functions. In particular, we measure the performance of both the architectures using various parameters. To the best of the authors' knowledge, this is the first attempt of its kind to apply CycleGAN and propose cross-domain architectures for the WHSP2SPCH conversion task. Statistically meaningful analysis of objective as well as subjective measures, is also presented.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

II. Proposed Cross-Domain Architectures

A. DiscoGAN with MMSE regularizer

Our goal is to learn mapping function between two domains, namely, whispered speech (W) and the normal speech (S). We denote the data distribution as $X_W \sim p_W$ and $X_S \sim p_S$, where X_W is the features of the whispered speech, and X_S is the features of the normal speech. Our model includes two mapping using generators G_{WS} and G_{SW} along with two discriminators D_W and D_S (as shown in Fig. 1). G_{WS} converts X_W into X_{WS} (converted features of the normal speech) such that X_{WS} is indistinguishable from the true samples X_S, and similarly for G_{SW}. Moreover, the discriminator D_S attempts to distinguish between X_S and X_{WS}. D_W performs an analogous operation for the X_W.

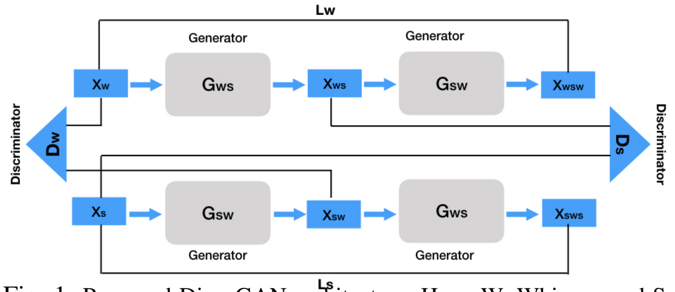

Fig. 1: Proposed DiscoGAN architecture. Here, W: Whisper, and S: Speech. After [16].

Our objective function contains regularized adversarial loss (Eq. 1), and the two reconstruction losses (Eq. 2). Here, an adversarial loss helps generator for matching the distribution in the target domain, and the regularization of this loss with MSE (as we can see in $\mathcal{L}_G$) helps in generating the samples that are corresponding to the given whispered speech utterances. Reconstruction losses help each generator to learn the mapping from its input domain to the output domain and discover relationships between them. These two objective functions are explored to encourage the one-to-one mapping between two domains. Since our task is to map the parameters of X_W to the parameters of X_S, we rely on the regularized adversarial objective function, which can be mathematically formulated as:

$$\begin{aligned}\mathcal{L}_G = &-\mathbb{E}_{X_S \sim p_S}[\log(D_W(G_{SW}(X_S)))] \\ &+ \frac{1}{2}\mathbb{E}_{X_W \sim p_W, X_S \sim p_S}[\log(X_W) - \log(G_{SW}(X_S))]^2,\end{aligned}$$

$$\begin{aligned}\mathcal{L}_D = &-\mathbb{E}_{X_W \sim p_W}[\log(D_W(X_W))] \\ &- \mathbb{E}_{X_S \sim p_S}[\log(1 - D_W(G_{SW}(X_S))],\end{aligned} \tag{1}$$

where $\mathbb{E}_{X_S \sim p_S}$ and $\mathbb{E}_{X_W \sim p_W}$ denotes the expectation over all the samples X_S and X_W coming from the distribution p_S and p_W, respectively. Here, G_{WS}, G_{SW}, D_W, and D_S must be jointly trained [32], including the two reconstruction losses, $\mathcal{L}_W$ and $\mathcal{L}_S$. This can be mathematically represented as:

$$\begin{aligned}\min_\theta \mathcal{L}_W &= \mathbb{E}[X_{WSW} - X_W]^2, \\ \min_\theta \mathcal{L}_S &= \mathbb{E}[X_{SWS} - X_S]^2.\end{aligned} \tag{2}$$

These reconstruction losses (given by eq. (2)) satisfy our requirement that G_{WS} and G_{SW} must be inverse of each other to the extent possible, i.e., for any X_W, $X_{WSW} = G_{SW}(G_{WS}(X_W))$ must be close to the X_W, and similarly for any X_S. Ideally, the equality of X_{WSW} and X_W (i.e., $X_{WSW} = X_W$) should hold. However, this is difficult to optimize [32]. For this reason, the distance between $\mathcal{L}_W$ and $\mathcal{L}_S$ is minimized by using the *MSE loss* [32], [36]. Hence, the total generator loss for G_{WS} can be defined as:

$$\mathcal{L}_{G_{WS}} = \mathcal{L}_W + \mathcal{L}_{G_S}, \tag{3}$$

where L_{G_S} can be defined in the form of eq. (1). The generator loss G_{SW} can also be defined in the similar way. Hence, total generator loss is $\mathcal{L}_{G_{WS}} + \mathcal{L}_{G_{SW}}$ and total discriminator loss is $\mathcal{L}_{D_W} + \mathcal{L}_{D_S}$, where $\mathcal{L}_{D_W}$ and $\mathcal{L}_{D_S}$ can also be defined in the form of eq. (1).

B. CycleGAN

Our model consists of two generators (i.e., G_{WS} and G_{SW}) and two discriminators (i.e., D_W and D_S) as illustrated in Fig. 2. Generator G_{WS} serves as a mapping function from X_W to X_S, and similarly for G_{SW}. The discriminators aims to distinguish between the real and generated distribution. For instance, D_W distinguish between X_S and X_{WS}, and D_S between X_W and X_{SW}.

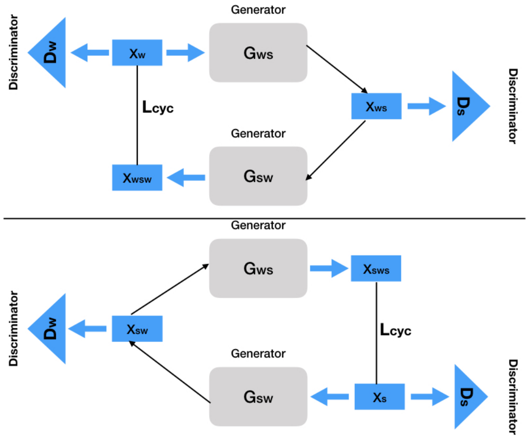

Fig. 2: Proposed CycleGAN architecture. Here, W: Whisper, and S: Speech. After [35].

We apply two types of loss functions defined as adversarial loss and cycle-consistent loss. In adversarial loss, we replace negative log-likelihood objective function by a least-squares loss for a better stability during training [35], [37]. The objective function for the adversarial loss for mapping G_{WS}, and corresponding discriminator D_S can be formulated as [35]:

$$\begin{aligned}\mathcal{L}_{G_{WS}} = &\mathbb{E}_{X_S \sim p_S}[(\log(D_S(X_S) - 1)^2] \\ &+ \mathbb{E}_{X_W \sim p_W}[\log(D_S(X_{WS}) - 1)^2].\end{aligned} \tag{4}$$

Similarly, we can define objective function for G_{SW}. Here, cycle-consistent loss ensure that an input X_W or X_S retain its original form after passing through two generators. The cycle-consistent loss function is analogous to the objective function of autoencoder, which minimizes the difference between the input and output to reconstruct the input from the output. We use metric function of the L_1 norm, and is defined as,

$$\begin{aligned}\mathcal{L}_{cyc} = \mathbb{E}_{X_W \sim P_W}[\|G_{SW}(G_{WS}(X_W)) - X_W\|_1] \\ + \mathbb{E}_{X_S \sim P_S}[\|G_{WS}(G_{SW}(X_S)) - X_S\|_1].\end{aligned} \quad (5)$$

These two losses incentivize the one-to-one mapping between two domains. The full objective function combines adversarial loss and cycle-consistent loss written as:

$$\mathcal{L}_{total} = \mathcal{L}_{G_{WS}} + \mathcal{L}_{G_{SW}} + \lambda\mathcal{L}_{cyc}. \quad (6)$$

where λ is hyper-parameter which controls the relative importance of cycle-consistent loss w.r.t. other losses. We have used $\lambda = 10$ during the experiments.

C. DiscoGAN vs. CycleGAN

Though the training method of both the architectures is the same, these architectures differ in terms of their loss functions. In CycleGAN, a least mean square loss (i.e., MSE) plays a role of adversarial loss, instead we have binary cross-entropy loss (BCE) as the adversarial loss in DiscoGAN. Moreover, the DiscoGAN has two different reconstruction losses which are regularized by MSE loss. However, CycleGAN has an only cycle-consistency loss which is regularized by the metric function of L_1 norm. The significance of the differences in loss functions in the context of WHSP2SPCH conversion is presented in the next section.

III. Experimental Results

A. Experimental Setup

In this paper, we have used the whispered TIMIT (wTIMIT) database [38]. In particular, we took one male and one female speaker's data for the development of WHSP2SPCH conversion systems. In total, 388 parallel utterances corresponding to the whispered and the normal speeches are taken for training and 35 utterances for testing. Each architecture is used to learn 1) mapping between the cepstral features corresponding to the whispered and the normal speech, and 2) the mapping between the converted cepstral features and the corresponding F_0 of the normal speech, which is followed by post-processing using a *sinc* interpolation smoothing in the voiced region.

In this paper, generators in GAN, DiscoGAN, and CycleGAN follow the identical architecture with the three hidden layers. Having a uniform architecture helps in analyzing the advantages of adversarial training equitably. Each hidden layer contains 512 neurons with Rectified Linear Unit (ReLU) activation, whereas the output layer has the linear activation function. The discriminators of the GAN, DiscoGAN, and CycleGAN also have three hidden layers, with ReLU activation function, whereas the output layer has sigmoid activation function. All the three models are trained for 80 epochs, using an effective batch size of 1000 frames as suggested in [39]. The parameters are optimized using Adam optimization, with a learning rate of 0.0001 [40]. The *40*-dimensional (dim) Mel Cepstral Coefficients (MCCs) (including the 0^{th} coefficient) are extracted from the whispered and normal speeches with 25 ms window and 5 ms frameshift. For analysis-synthesis, we have used AHOCODER [41].

B. 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 [42]:

$$MCD \text{ [in dB]} = \frac{10}{ln10}\sqrt{2\sum_{i=1}^{40}(m_i^t - m_i^c)^2}\ , \quad (7)$$

where m_i^t and m_i^c are the i^{th} MCCs of the reference, and converted signal. In particular, m_i^t and m_i^c are the i^{th} MCCs of the reference neutral speech and the converted neutral speech in the case of 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.

TABLE I: MCD analysis of the different systems. Here, % in the bracket indicates relative reduction in the MCD w.r.t the baseline

GAN Architectures	Male Speaker	Female Speaker
Baseline: GAN	7.04	8.12
CycleGAN	**6.39** **(9.23%)**	6.72 (17.21%)
DiscoGAN	6.6 (6.25%)	**6.63** **(18.35%)**

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

GAN Architectures	Male Speaker	Female Speaker
Baseline: GAN	5.23	4.56
CycleGAN	4.08 (21.99%)	5.88 (-28.94%)
DiscoGAN	**2.54** **(51.43%)**	**4.52** **(0.9%)**

To measure the RMSE of $\log(F_0)$, the actual reference speech and the converted speech signals, are time-aligned using the Dynamic Time Warpping (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 absence of voicing) [43]. RMSE of the $\log(F_0)$ is given by:

$$RMSE(\log(F_0)) = \sqrt{\sum_{i=1}^{k}[\log(F_{0_i}^t) - \log(F_{0_i}^c)]^2}, \quad (8)$$

where k is the total number of voiced-voiced pairs after the alignment, and ${F_0}^t$ 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.

The effectiveness of the CycleGAN and DiscoGAN can be clearly seen for the WHSP2SPCH conversion system in objective results. As shown in Table I and Table II, we can see the relative % reduction in MCD and RMSE of $\log(F_0)$ for the male and female speakers over the baseline, respectively. The noteworthy performance of both the proposed architectures is due to their ability to encapsulate the cross-domain relationships efficiently compared to GAN [32], [35], [44]. We also compare the results of CycleGAN and DiscoGAN, where the MCD results of both the architectures are comparable. However, the DiscoGAN outperforms the CycleGAN in terms of RMSE of the $\log(F_0)$, and gets an overall 37.75% and 23.13% reduction in RMSE of the $\log(F_0)$ of male and female speakers, respectively. An example of the generated F_0 contour using the various developed systems are shown in Fig. 3. We used the same DNN-based architectures for the voice-unvoiced decision, which can be seen in the Fig. 3.

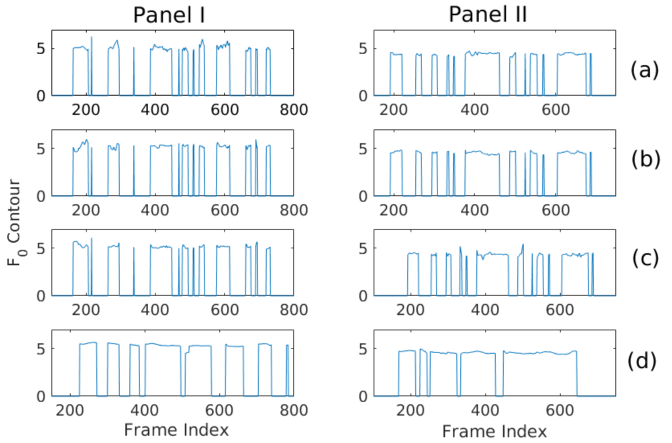

Fig. 3: $\log(F_0)$ predicted using the (a) GAN, (b) CycleGAN, (c) DiscoGAN, and (d) corresponding natural speech signal for Panel I: female, and Panel II: male speakers.

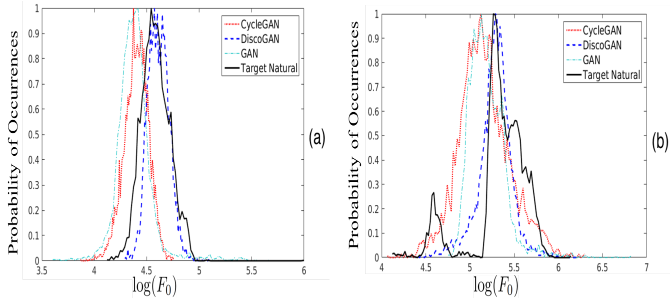

Fig. 4: Histogram of generated $\log(F_0)$ using different techniques for (a) male, and (b) female speaker along with the histogram of $\log(F_0)$ of natural speech signal.

The distribution of the predicted $\log(F_0)$ is presented in Fig. 4 for evaluating the performance of the WHSP2SPCH conversion systems. It is very clear from Fig. 4 that the distribution of the generated F_0, using the proposed DiscoGAN architecture, closely follows the distribution of the F_0 corresponding to the natural speech. This finding is in line with key objective of GAN architectures [33].

C. Subjective Evaluation

Comparative subjective analysis test, namely, ABX has been taken for the subjective evaluations. Total 41 subjects (21 females and 20 males between 18 to 30 years of age and with no known hearing impairments) took part in the subjective test. Here, we randomly played same utterances from two different systems and asked subjects to decide which one is more better in terms of naturalness. Results of the ABX tests obtained from the total 504 samples are shown in Fig. 5. We can clearly see that the proposed DiscoGAN and CycleGAN are 24.39% and 87.1% times more preferred over the GAN by the subjects. Furthermore, we observed that the CycleGAN is 53.7% times more preferred than the DiscoGAN.

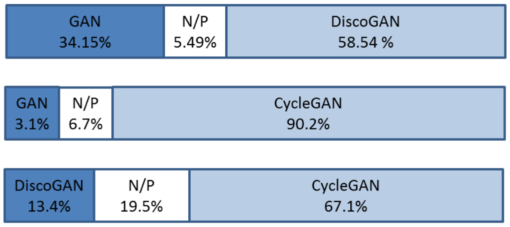

Fig. 5: ABX test analysis for the various developed systems.

Interestingly, we found that DiscoGAN performs better in objective results, and CycleGAN performs better in subjective results. The key reasons for this recline in their differences lies in loss functions. We know that MCD and RMSE of $\log(F_0)$ are Euclidean distances and encapsulate the characteristics of MSE loss, as shown in eq. (7) and eq. (8). The goal of our proposed architectures is to update weights and biases via MSE loss. This training method using MSE loss as reconstruction loss in DiscoGAN helps to reduce the difference between the real and the generated data distributions, which leads to desirable results in our objective functions. The effect of outliers is exponential in L_2 norm (here, MSE loss) [45]. As discussed in [46], [47], outliers degrade the speech quality in terms of intelligibility and naturalness. However, L_1 norm is less susceptible to the outliers [45], which leads to better subjective results using CycleGAN since we are using L_1 norm as cycle consistency loss. Another possible reason is that L_2 norm has the unique possible shortest path/solution in Euclidean space which helps to minimize MSE loss efficiently [45], yields better objective results. On contradictory, L_1 norm has multiple path/solutions due to absolute value (i.e., mode operation [45]), which helps to capture different style of speaking by different speakers more efficiently and hence, yielding better results in the subjective evaluation.

IV. Summary and Conclusions

In this paper, cross-domain CycleGAN and DiscoGAN architectures have been proposed for finding the cross-domain relationships between the whisper and normal speeches. The proposed cross-domain architectures perform better compared to the baseline GAN in both the objective as well as the subjective evaluations. Furthermore, it has been observed that the DiscoGAN performs better in objective results, whereas the CycleGAN performs better in the subjective results. This is primarily due to the differences in the loss functions of both the architectures. In addition, we found that the distributions of the generated F_0 obtained using the proposed architectures closely follow the distribution of the F_0 corresponding to the natural speech signal. However, there is still the room for improvement in the quality of the converted voices. In the future, we plan to explore high-quality vocoder, namely, WaveNet for further improvement in voice quality.

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] 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.

[7] 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.

[8] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, no. 7, pp. 515–520, 2002.

[9] I. 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, p. 12, 2015.

[10] I. V. McLoughlin, J. Li, and Y. Song, "Reconstruction of continuous voiced speech from whispers," in *INTERSPEECH*, Lyon, France, 2013, pp. 1022–1026.

[11] T. Toda, M. Nakagiri, and K. 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.

[12] V.-A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Multimodal HMM-based NAM-to-speech conversion," in *INTERSPEECH*, Brighton, United Kingdom (UK), 2009, pp. 656–659.

[13] V. A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Improvement to a NAM-captured whisper-to-speech system," *Speech Communication*, vol. 52, no. 4, pp. 314–326, 2010.

[14] G. N. Meenakshi and P. K. Ghosh, "Whispered speech-to-neutral speech conversion using bidirectional LSTMs," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 491–495.

[15] 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.

[16] 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.

[17] 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.

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

[19] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[20] 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.

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

[22] V. C. Tartter, "What is in a whisper," *JASA*, vol. 86, no. 5, pp. 1678–1683, 1989.

[23] K. N. Stevens, *Acoustic Phonetics*. MIT Press, 2000.

[24] 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*, Brighton, UK, May,2019.

[25] G. B. Remijn *et al.*, "A near-infrared spectroscopy study on cortical hemodynamic responses to normal and whispered speech in 3-to 7-year-old children," *Journal of Speech, Language, and Hearing Research (JSLHR)*, vol. 60, no. 2, pp. 465–470, 2017.

[26] Y. Tajiri, K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Non-audible murmur enhancement based on statistical conversion using air-and body-conductive microphones in noisy environments," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2769–2773.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *ICML*, Sydney, Australia, 2017, pp. 1857–1865.

[33] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, Jan 2018.

[34] K. Oyamada, H. Kameoka, T. Kaneko, H. Ando, K. Hiramatsu, and K. Kashino, "Non-native speech conversion with consistency-aware recursive network and generative adversarial network," in *APSIPA ASC*, Kuala Lumpur, Malaysia, 2017, pp. 182–188.

[35] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, Venice, Italy, 2017, pp. 1–18.

[36] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[37] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017, pp. 2794–2802.

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

[39] N. Shah, N. J. Shah, and H. A. Patil, "Effectiveness of generative adversarial network for non-audible murmur-to-whisper speech conversion," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 3157–3161.

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] E. Kreyszig, *Introductory Functional Analysis with Applications*. wiley New York, 1^{st} (Eds.), 1989, vol. 81.

[46] S. V. Rao, N. J. Shah, and H. A. Patil, "Novel pre-processing using outlier removal in voice conversion," in *Speech Synthesis Workshop (SSW)*, Sunnyvale, CA, USA, 2016, pp. 134–139.

[47] N. J. Shah and H. A. Patil, "A novel approach to remove outliers for parallel voice conversion," *Computer Speech & Language*, vol. 58, pp. 127–152, 2019.