INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Whisper Augmented End-to-End/Hybrid Speech Recognition System - CycleGAN Approach

Prithvi R. R. Gudepu, Gowtham P. Vadisetti, Abhishek Niranjan, Kinnera Saranu, Raghava Sarma, M Ali Basha Shaik, Periyasamy Paramasivam

Samsung Research Institute India - Bangalore
{p.gudepu, gowtham.p, a.niranjan, kinnera.sar, raghava70.a}@samsung.com, {m.shaik, periyasamy.p}@samsung.com

Abstract

Automatic speech recognition (ASR) systems are known to perform poorly under whispered speech conditions. One of the primary reasons is the lack of large annotated whisper corpora. To address this challenge, we propose data augmentation with synthetic whisper corpus generated from normal speech using Cycle-Consistent Generative Adversarial Network (CycleGAN). We train CycleGAN model with a limited corpus of parallel whispered and normal speech, aligned using Dynamic Time Warping (DTW). The model learns frame-wise mapping from feature vectors of normal speech to those of whisper. We then augment ASR systems with the generated synthetic whisper corpus. In this paper, we validate our proposed approach using state-of-the-art end-to-end (E2E) and hybrid ASR systems trained on publicly available Librispeech, wTIMIT and internally recorded far-field corpora. We achieved 23% relative reduction in word error rate (WER) compared to baseline on whisper test sets. In addition, we also achieved WER reductions on Librispeech and far-field test sets.

Index Terms: ASR, CycleGAN, speech-to-whisper, DTW, E2E

1. Introduction

Virtual Assistants have become part of our daily life. They help in getting work done with intuitive and direct voice commands making the tedious touch based interface redundant. Having seamless usage requires the assistants, especially speech recognition systems they contain, to work in many challenging scenarios including when the user whispers. Whisper is particularly useful when the user needs to interact with the assistant while in a meeting or at a public place, to keep the interaction private and subtle. In addition, whisper is useful in home environment to avoid disturbing others while conversing with assistants.

Regular Automatic Speech Recognition (ASR) systems built on frameworks such as RETURNN [1], EESEN [2], and DeepSpeech [3] and trained with normal speech might not perform well on whisper due to significant differences between acoustic characteristics of normal and whispered speech. Normal or phonated speech is generated by the vibration of vocal folds and resonance of the vocal cord that releases air in short pulses. This vibration is observed as fundamental frequency F0 but is absent in the case of whispered speech as the vocal folds are held open to allow the air to pass through. Additionally, the difference in the sound production mechanism leads to shifting of formants towards higher frequencies and broadening of formant ranges thereby spreading the spectrum content. The important information (i.e. speaker dependent information, change of signal characteristics, etc.) distinguishing different phones is lost and an ASR system that has not seen whispered speech in its training has a lot more confusion in distinguishing phones leading to a sharp drop in performance. These complexities in general and its similarities with acoustic characteristics of noise make whispered speech recognition a challenging task.

Data augmentation is a proven technique for improving the performance of speech recognition, where huge amounts of speech is used in general [4, 5, 6, 7]. However due to the dearth of whisper data compared to speech corpora, whisper recognition remains a challenging task. So, we propose a novel synthetic whisper data augmentation technique based on CycleGAN [8].

Generative Adversarial Network (GAN) [9] is one of the widely known frameworks among generative modelling approaches. It learns the training data distribution under adversarial conditions, to generate target samples having same statistics. It comprises of a generator that learns distributions and a discriminator that learns to differentiate real samples from generated ones. They are particularly used for style or domain transfer tasks [8,10,11,12,13,14]. CycleGANs have been successful in computer vision tasks such as image-to-image translation [8], speech related tasks such as emotion style transfer [15], voice transformation on impaired speech [16], and voice conversion [17].

In the literature, Denoising Autoencoders have been used for generating pseudo-whisper [18] and simple DNNs have been used for the purpose of generating whispered speech from phonated speech [19], but CycleGANs give an edge due to their adversarial architecture and stabilization of GAN training due to their cycle consistency. Though CycleGANs can be used with unpaired data, better performance is achieved with paired data and for pairing whisper and normal speech, techniques such as Dynamic Time Warping (DTW) can be used.

To the best of our knowledge, this is the first attempt to use CycleGAN's data augmentation mechanism for improving ASR performance, both on whisper as well as far field speech. In section 2, we introduce CycleGAN architecture and present how we use it to generate synthetic whisper data from a large speech corpus using WORLD vocoder [21] for speech synthesis. In section 3, we present RWTH's RETURNN end-to-end (E2E) ASR [1] and a Hybrid ASR system [22], and analyze the quality of the generated synthetic whisper data by training them with different combinations of synthetic whisper, natural whisper, and normal speech. In section 4, we report the results and show that augmenting using synthetic data boosts the performance of baseline models. We also show that in normal test cases, the

http://dx.doi.org/10.21437/Interspeech.2020-2639

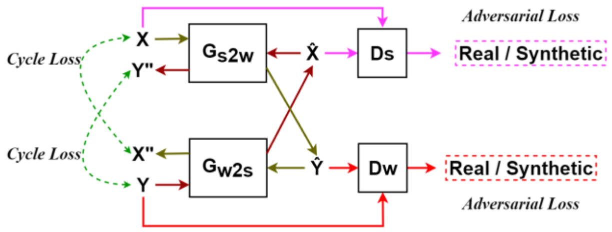

Figure 1: *CycleGAN consisting of mapping functions G_{s2w} , G_{w2s} and discriminators D_s , D_w , respectively.*

performance is maintained or improved considerably when augmented with synthetic whisper data.

2. Proposed Method

2.1. Cycle-Consistent Generative Adversarial Network

CycleGAN [8] consists of two generators G_{s2w} , G_{w2s} and two discriminators D_s, D_w. G_{s2w} maps speech (X) to whisper (Y) while G_{w2s} maps whisper to speech. $G_{s2w}: X \to Y$ and $G_{w2s}: Y \to X$, both these functions are bijections and inverses of each other. $\hat{X} = G_{w2s}(Y)$, $\hat{Y} = G_{s2w}(X)$ are the generated speech and whisper samples respectively, as shown in Figure 1. $X" = G_{w2s}(\hat{Y})$, $Y" = G_{s2w}(\hat{X})$ are the reconstructed speech and whisper samples converted from their target domain. After passing through each of these generators once, they tend to form their original feature vector, hence ensuring cycle consistency. The discriminators try to differentiate between real and fake samples in their respective domain.

The loss function has three components: The *adversarial loss*($\mathcal{L}_{gan}$), shown in (1), tends to bring the real and converted sample i.e. X and $\hat{X}$ or Y and $\hat{Y}$ close as possible using the mean squared error. The *cycle-consistency loss*($\mathcal{L}_{cyc}$), shown in (2), ensures that X or Y tends to retain its original feature vector after passing through the two generators, using L1 loss. The *identity loss*($\mathcal{L}_{id}$), shown in (3), ensures that X or Y when belongs to target domain is not transformed, using L1 loss.

$$\begin{aligned}\mathcal{L}_{gan}(G_{s2w}, D_w, X, Y) = {} & \mathbb{E}_{y \sim P_{data}(Y)}[(D_w(y))^2] \\ & + \mathbb{E}_{x \sim P_{data}(X)}[(1 - D_w(G_{s2w}(x))^2]\end{aligned} \quad (1)$$

$$\begin{aligned}\mathcal{L}_{cyc}(G_{s2w}, G_{w2s}, X, Y) = {} & \mathbb{E}_{x \sim P_{data}(X)}\big[\parallel G_{w2s}\big(G_{s2w}(x)\big) - x \parallel_1\big] \\ & + \mathbb{E}_{y \sim P_{data}(Y)}\big[\parallel G_{s2w}\big(G_{w2s}(y)\big) - y \parallel_1\big]\end{aligned} \quad (2)$$

$$\begin{aligned}\mathcal{L}_{id}(G_{s2w}, G_{w2s}, X, Y) = {} & \mathbb{E}_{x \sim P_{data}(X)}[\parallel G_{w2s}(x) - x \parallel_1] \\ & + \mathbb{E}_{y \sim P_{data}(Y)}[\parallel G_{s2w}(y) - y \parallel_1]\end{aligned} \quad (3)$$

The model weights are learnt as the generators and discriminators compete by applying the minimax defined as

$$G_{s2w}{}^*, G_{w2s}{}^* = arg\ min_G\ max_D\ \mathcal{L}(G_{s2w}, G_{w2s}, D_w, D_s) \quad (4)$$

where,

$$\begin{aligned}\mathcal{L}(G_{s2w}, G_{w2s}, D_w, D_s) = {} & \mathcal{L}_{gan}(G_{s2w}, D_w, X, Y) + \mathcal{L}_{gan}(G_{w2s}, D_s, X, Y) \\ & + \lambda_{cyc} * \mathcal{L}_{cyc}(G_{s2w}, G_{w2s}, X, Y) \\ & + \lambda_{id} * \mathcal{L}_{id}(G_{s2w}, G_{w2s}, X, Y)\end{aligned} \quad (5)$$

using the back propagation algorithm. The generators try to fool the discriminators, and the discriminators try to identify the fake samples and in this adversarial setup all these networks become proficient at their respective tasks.

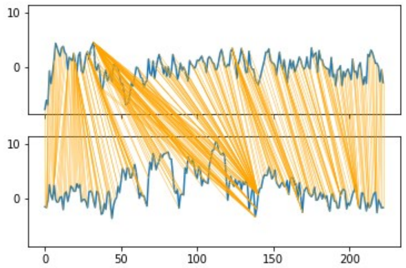

Figure 2: *DTW optimal mapping between speech and whisper frames(shown for only 1 dimension for better visualization).*

2.2. DTW

Dynamic Time Warping (DTW) is used to align feature vectors of corresponding normal and whisper audios in the parallel data used for training the CycleGAN model. DTW is a signal alignment algorithm, which matches two temporal sequences with a monotonically increasing optimal warping path satisfying boundary conditions. Let, X and Y be N-dimensional frame wise sequence of speech and whisper feature vectors, respectively as shown below.

$$\begin{aligned}X &= x_1, x_2, \dots, x_n \\ Y &= y_1, y_2, \dots, y_n \\ x_i &= [x_{i1}, x_{i2}, \dots, x_{iN}] \\ y_j &= [y_{j1}, y_{j2}, \dots, y_{jN}]\end{aligned} \quad (6)$$

where, n and m are the number of frames or feature vectors in speech and whisper. An optimal path $W = w_1, w_2, \dots, w_K$ is computed with the condition $max(n, m) \leq K \leq n + m$, where $w_k = (i, j)$, $w_{k+1} = (i', j')$ such that $i \leq i' \leq i + 1$, $j \leq j' \leq j + 1$. i and j are monotonically increasing indices in speech and whisper frames respectively. The warp path as shown in Figure 2 is such that the following Euclidean distance is minimized.

$$Dist(W) = \sum_{k=1}^{k=K} Dist(w_{ki}, w_{kj}) \quad (7)$$

3. Experimental Setup

3.1. ASR systems

To make sure that the proposed approach has wider applicability and is ASR agnostic, we evaluate it with both E2E ASR and Hybrid ASR systems as described below

E2E ASR: We choose RWTH Aachen's open source state of the art RETURNN [1] toolkit, which is based on E2E encoder-attention-decoder architecture. Log-filterbank energy (LFBE) features are used. We use the trained RETURNN models without pre-trained LSTM language model, presented in [23], in recognition step.

Hybrid ASR: It comprises a separate acoustic and an n-gram language model. We use Kaldi [24] for GMM-HMM training and tensorflow for training an LSTM acoustic model consisting of 4 hidden layers with sizes 1100, 990, 880 and 770. We follow standard Kaldi recipe for WSJ corpus to train LDA-MLLT GMM-HMM. We use this model to generate senone alignment

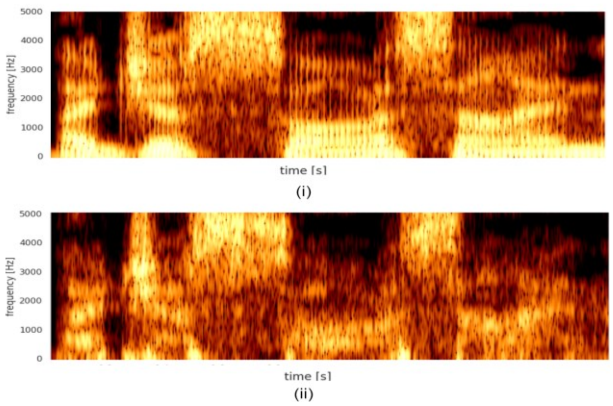

Figure 3: *Spectrogram of normal speech(i) and synthetic whisper generated using CycleGAN(ii).*

for training the acoustic model used as baseline. The 39-dim features used in the experiments include the 13-dim static Mel-frequency cepstral coefficient (MFCC) (with C0 replaced with energy) and its first and second derivatives. The features were pre-processed with the cepstral mean normalization (CMN) algorithm. Decoding is done using a 3-gram LM, which is further rescored using a 6-gram language model having vocabulary size of one million words. The training procedure of the ASR system is similar to the one described in [22].

3.2. Datasets

Librispeech. [25] We use this data set for training baseline ASR systems and as the source for generating synthetic whisper data set. It contains 960 hours of speech sampled at 16 kHz. The training portion of the corpus has three subsets, approximately with sizes 100, 360, and 500 hours. The entire 960 hours data is used for training RETURNN ASR while train-clean-100 and train-clean-360 are used for generating synthetic whisper data.
wTIMIT. [26] It consists of 450 phonetically balanced utterances in both normal and whispered speech with two accents: Singaporean-English and North American with 24 and 28 speakers, respectively [26]. It is recorded at 44.1kHz and is gender balanced. It consists of around 26 hours of parallel normal and whispered data.
Internally Recorded Trainset (TR1). It consists of 450 utterances both in normal and whispered speech internally recorded by 200 speakers under clean conditions. It is recorded at 44.1kHz and is gender balanced. It consists of 180 hours of parallel normal and whispered near-field data.
Internally Recorded Testset (TS1). It contains a total of 2500 internally recorded utterances of 20 speakers in both normal and whispered speech under normal office environment sampled at 16kHz. We further add different noise profiles to them using Kaldi wav-reverberate [28] mechanism taking the total to 10000 utterances.

Internally Recorded Far-field Testset (TS2). It contains 5200 normal speech utterances recorded by 20 speakers at 1m and 3m distances under clean conditions. The audio files are recorded with sampling rate of 16kHz and the dataset is gender balanced. Additional test cases are generated by mixing background noises such as babble, kitchen and TV using Kaldi wav-reverberate mechanism taking the total number of test cases to 20800.
All internal data sets are recorded in North American Accent.

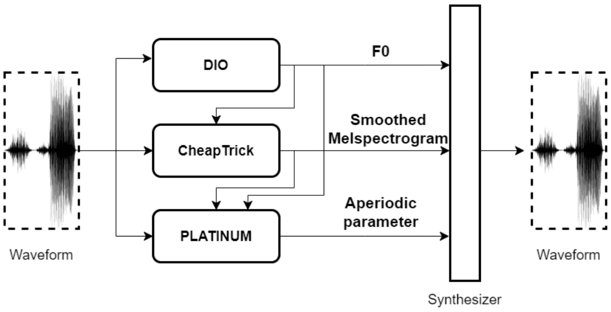

Figure 4: *Structure of World Vocoder consisting of three algorithms DIO, CheapTrick and PLATINUM to estimate three parameters and a synthesizer to generate waveform.*

3.3. WORLD Vocoder

To obtain audio from whisper feature vectors synthesized with CycleGAN, we use WORLD [21] vocoder, shown in Figure 4. It consists of three analysis algorithms for determining fundamental frequency F0, spectral envelope, and aperiodic parameters and a synthesis algorithm to generate speech signal using these three parameters. F0 is calculated by an estimation algorithm named Distributed Inline-filter Operation (DIO) [27]. The spectral envelope consists of 24-dimensional smoothed spectrogram features, estimated by algorithm called CheapTrick [28]. Aperiodic parameter is extracted using algorithm called PLATINUM [29]. Since whisper does not contain pitch, F0 is made zero and aperiodic parameter is taken as 513-dimensional unit vector as the source of whisper is aperiodic due to lack of periodic airflow. Whisper signal is generated using these parameters using a synthesizer [21] module.

3.4. CycleGAN Setup and Training

The generators and discriminators are implemented as standard feed-forward DNNs each having three hidden layers with 512 neurons each and Rectified Linear Units (ReLUs) as activation function. The model is trained for 300 epochs with learning rate set to 0.0001 and weight decay to 1e^{-5}. For training we use Adam optimization [30] and a batch size of 1000 implemented in PyTorch version 1.2.0.

The relative weights of *identity* and *cycle-consistency loss,* λ_{id} and λ_{cyc}, are set to 5 and 10 respectively. CycleGAN is trained using WORLD's [21] 24-dimensional smoothed melspectrogram features extracted from 30 hours of TR1. We use the trained model to convert the smoothed mel-spectrogram features extracted from 460 hours of Librispeech consisting of train-clean-100 and train-clean-360 datasets into synthetic whisper features. F0 is taken to be zero and aperiodic parameter is taken to be 513-dimensional unit vector for all frames of synthetic whisper and synthesizer is used to generate the whisper signals forming the synthetic whispered dataset (TR2). Spectrogram of a sample normal speech signal and its converted synthetic whisper speech signal is shown in Figure 3.

3.5. ASR Experiments

E2E RETURNN ASR model (b0) is trained with 960 hours of Librispeech data using 40-dimensional LFBE features and taken as baseline model for E2E ASR experiments. Next, we train an augmented synthetic whisper model (s1) with training data of baseline model b0 combined with 460 hours of synthetic whispered data generated from Librispeech dataset (TR2).

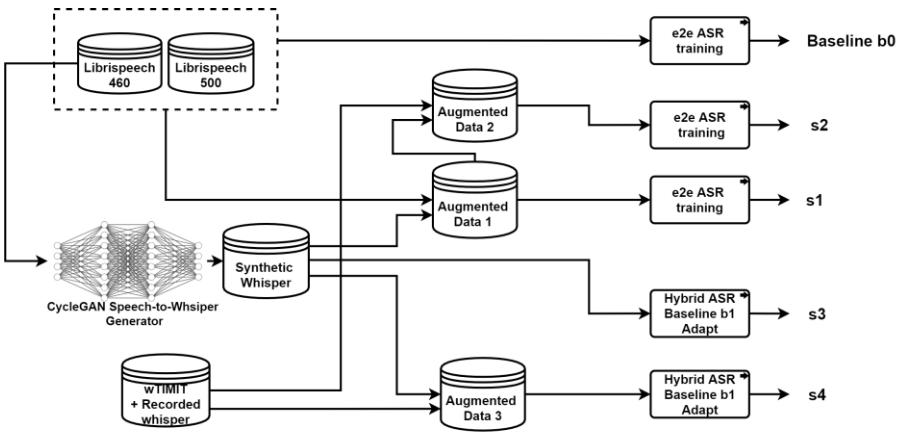

Figure 5: *Block diagram of the ASR experiments*

Model s2 is trained on training data of model s1 with additional wTIMIT (normal and whisper) and internally recorded whisper (TR1) train set.

We train another baseline model, which is Hybrid ASR (b1), as described in section 3.1. The next model (s3) is trained with training data of model (b1) added with synthetic whispered data (TR2). The model (s4) is trained with training data of (s3) combined with wTIMIT (normal and whisper) and internally recorded whisper (TR1) train set. The results of all the models with testing on Librispeech and wTIMIT are presented in Table 1. The performance of Hybrid ASR model and its derivatives with internally recorded near-field (TS1) and far-field (TS2) test sets are shown in Table 2. The block diagram of the experiments is shown in Figure 5.

Table 1: *WER (%) comparison of end-to-end (b0,s1,s2) and hybrid(b1,s3,s4) models*

Model		**LibriSpeech**				**wTIMIT Test**	
Id	**Sys**	dev-cln	dev-oth	test-cln	test-oth	norm	whsp
b0	Baseline	6.5	16.9	6.2	18.0	12.9	37.1
s1	+Wsyn	6.5	17.1	6.3	18.1	12.1	32.1
s2	+Wnat	6.1	15.9	5.9	16.9	9.5	29.4
b1	Baseline	6.8	16.3	7.1	16.2	19.2	44.3
s3	+Wsyn	6.8	16.3	7.3	16.4	18.3	40.9
s4	+Wnat	6.9	16.4	7.3	16.4	18.2	35.0

4. Results

Table 1 presents the results obtained for test sets of LibriSpeech and wTIMIT. The models s1 (E2E) and s3 (Hybrid), trained with synthetic whisper (TR2) in addition to their respective baseline models, show relative WER reductions of 13.5% and 7.6%, respectively, on wTIMIT whisper test set, while on normal test sets, both show minimal increase for LibriSpeech and slight decrease for wTIMIT. Table 2 presents the results obtained for test sets (TS1) and (TS2). The model s3 (Hybrid) shows relative WER reduction of 23%, 20% and 16% on clean whisper, noisy whisper and normal speech test cases of the test set (TS1) respectively. It achieves relative WER reduction of 26% and 15% respectively on 5 meter and 3 meter test cases of far field normal speech test set (TS2). The improvements on (TS1) and (TS2) test sets are significant considering the fact that they have more number of utterances and speakers, have wider noise profile added, and match the application environment.

The models s2 (E2E) and s4 (Hybrid) are trained with an additional corpus of wTIMIT (normal and whisper) and recorded whisper data (TR1), respectively. As shown in Table 1, they both have further reduced whisper WER by 8% and 14% respectively in wTIMIT whisper without degrading performance for normal speech. We can infer that there is more scope for improvement with CycleGAN based model that can be explored.

Table 2: *WER(%) comparison of Hybrid models on near-field(TS1) and far-field(TS2) testsets*

Model	**Near-field**				**Far-field**	
Id	whsp (clean)	whsp (noisy)	norm (clean)	norm (noisy)	1m	3m
b1	27.9	30.9	7.2	7.4	6.6	10.9
s3	21.5	24.8	6.2	6.1	5.6	8.0
s4	18.6	21.9	5.4	6.1	5.6	7.7

5. Conclusions and Future Work

We have proposed a novel data augmentation method using CycleGAN for whispered speech recognition. We achieved noticeable WER reduction on whispered test set while maintaining its performance with near-field speech and improved performance with far-field speech. Similar gains in performance is observed irrespective of the features (MFCC or LFBE) used. The proposed approach is promising in expanding the reach of whisper recognition, particularly for low resource languages under limited or sparse data conditions. The experiments with natural recorded whisper show that the CycleGAN approach has still scope for improvement, which we will be pursuing as part of our future work. We also observe that the CycleGAN model may have less language dependency and we plan to build unified models. We plan to release CycleGAN models as open source in the near future.

6. Acknowledgements

This work was funded by the AI R&D Group and Technology Strategy Groups in Samsung Electronics, Korea. We especially thank Vice President Kyoung-Gu Woo and Dr. Sky Jung for their support and guidance. We also thank Nirmesh J. Shah for his contributions in the initial phase of the research.

7. References

[1] Doetsch, Patrick, Albert Zeyer, Paul Voigtlaender, Ilia Kulikov, Ralf Schlüter, and Hermann Ney. "RETURNN: The RWTH extensible training framework for universal recurrent neural networks." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5345-5349. IEEE, 2017.

[2] Miao, Yajie, Mohammad Gowayyed, and Florian Metze. "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding." In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 167-174. IEEE, 2015.

[3] Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger et al. "Deep speech: Scaling up end-to-end speech recognition." *arXiv preprint arXiv:1412.5567* (2014).

[4] Park, Daniel S., William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. "Specaugment: A simple data augmentation method for automatic speech recognition." *arXiv preprint arXiv:1904.08779* (2019).

[5] Thai, Bao, Robert Jimerson, Dominic Arcoraci, Emily Prud'hommeaux, and Raymond Ptucha. "Synthetic Data Augmentation for Improving Low-Resource ASR." In *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, pp. 1-9. IEEE, 2019.

[6] Ramirez, Jose Manuel, Ana Montalvo, and Jose Ramon Calvo. "A Survey of the Effects of Data Augmentation for Automatic Speech Recognition Systems." In *Iberoamerican Congress on Pattern Recognition*, pp. 669-678. Springer, Cham, 2019.

[7] Hayashi, Tomoki, Shinji Watanabe, Yu Zhang, Tomoki Toda, Takaaki Hori, Ramon Astudillo, and Kazuya Takeda. "Back-translation-style data augmentation for end-to-end ASR." In 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 426-433. IEEE, 2018.

[8] Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 2223-2232. 2017.

[9] Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial networks." *Advances in neural information processing systems* 3, no. 06 (2014).

[10] Kim, Taeksoo, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. "Learning to discover cross-domain relations with generative adversarial networks." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1857-1865. JMLR. org, 2017.

[11] Shrivastava, Ashish, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. "Learning from simulated and unsupervised images through adversarial training." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2107-2116. 2017.

[12] Taigman, Yaniv, Adam Polyak, and Lior Wolf. "Unsupervised cross-domain image generation." *arXiv preprint arXiv:1611.02200* (2016).

[13] Liu, Ming-Yu, Thomas Breuel, and Jan Kautz. "Unsupervised image-to-image translation networks." In *Advances in neural information processing systems*, pp. 700-708. 2017.

[14] Yi, Zili, Hao Zhang, Ping Tan, and Minglun Gong. "Dualgan: Unsupervised dual learning for image-to-image translation." In *Proceedings of the IEEE international conference on computer vision*, pp. 2849-2857. 2017.

[15] Bao, Fang, Michael Neumann, and Ngoc Thang Vu. "CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition." *Manuscript submitted for publication* (2019): 35-37.

[16] Chen, Li-Wei, Hung-Yi Lee, and Yu Tsao. "Generative adversarial networks for unpaired voice transformation on impaired speech." *arXiv preprint arXiv:1810.12656* (2018).

[17] Chen, Ling-Hui, 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* 22, no. 12 (2014): 1859-1872.

[18] Ghaffarzadegan, Shabnam, Hynek Bořil, and John HL Hansen. "Generative modeling of pseudo-whisper for robust whispered speech recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, no. 10 (2016): 1705-1720.

[19] Cotescu, Marius, Thomas Drugman, Goeric Huybrechts, Jaime Lorenzo-Trueba, and Alexis Moinet. "Voice Conversion for Whispered Speech Synthesis." *IEEE Signal Processing Letters* (2019).

[20] Berndt, Donald J., and James Clifford. "Using dynamic time warping to find patterns in time series." In *KDD workshop*, vol. 10, no. 16, pp. 359-370. 1994.

[21] Morise, Masanori, Fumiya Yokomori, and Kenji Ozawa. "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications." *IEICE TRANSACTIONS on Information and Systems* 99, no. 7 (2016): 1877-1884.

[22] Kim, JC Jungsuk, and I. Lane. "Hydra a hybrid cpu/gpu speech recognition engine for real-time lvcsr." In *GPU Technology Conference*. 2013.

[23] Zeyer, Albert, Kazuki Irie, Ralf Schlüter and Hermann Ney. “Improved training of E2E attention models for speech recognition.” *INTERSPEECH* (2018).

[24] Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann et al. "The Kaldi speech recognition toolkit." In *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[25] Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. "Librispeech: an asr corpus based on public domain audio books." In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206-5210. IEEE, 2015.

[26] Lim, Boon Pang. "Computational differences between whispered and non-whispered speech." PhD diss., University of Illinois at Urbana-Champaign, 2011.

[27] Morise, Masanori, Hideki Kawahara, and Haruhiro Katayose. "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech." In *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society, 2009.

[28] Morise, Masanori. "CheapTrick, a spectral envelope estimator for high-quality speech synthesis." *Speech Communication* 67 (2015): 1-7.

[29] Morise, Masanori. "Platinum: A method to extract excitation signals for voice synthesis system." *Acoustical Science and Technology* 33, no. 2 (2012): 123-125.

[30] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).