Data Augmentation for ASR using CycleGAN-VC

Stanford CS224N Custom Project
External collaborator: Claire Pajot (CS224S)
Mentor: Angelica Sun

Sofian Zalouk †
Stanford University
szalouk@stanford.edu

Hikaru Hotta †
Stanford University
hhotta@stanford.edu

Claire Pajot †
Stanford University
cpajot@stanford.edu

Abstract

There is a significant performance gap in ASR systems between black and white speakers, which is attributed to insufficient audio data from black speakers available for models to train on. We aim to close this gap by using a CycleGAN based voice converter to generate African American Vernacular English utterances from generic American English utterances as a data augmentation strategy. By using a two-step adversarial loss and a self-supervised frame filling task, we were able to noticeably improve the qualitative performance of our CycleGAN based voice conversion pipeline. In spite of this, we could not establish the method of CycleGAN based voice conversion as a reliable method for data augmentation. While this project was challenging, it was especially rewarding to conduct this line of research which has the ultimate goal of ensuring that marginalized voices are heard. All code is publicly available at our Github repository.

1 Introduction

Millions of people around the world use Automated Speech Recognition (ASR) systems to transcribe their speech through applications like virtual assistants on mobile devices, captioning technology, and speech interfaces in cars. However, as with other applications of machine learning, there is increasing concern that speech recognition suffers from harmful biases. A recent paper by Koenecke et al. (2020) showed that commercial ASR systems from companies like Amazon, Google, and Apple made twice as many errors transcribing audio from black speakers as audio from white speakers [1]. The authors hypothesize that this disparity can be traced back to a performance gap in acoustic models in ASR systems due to insufficient audio data from black speakers when training the models.

Koenecke et al. suggest that ASR systems can be made more inclusive through better data collection for nonstandard varieties of English, such as regional accents or nonnative-English accents. However, ensuring that this is achieved for all subgroups of the population is challenging. We evaluate speech recognition performance on a small speech dataset from an underrepresented subgroup, in our case African American Vernacular English (AAVE) speakers, when augmented using a CycleGAN-based voice converter combined with audio from a dataset of more commonly represented speakers, in our case white Californian speakers. Thus, we propose a computational method to overcome some of the challenges of limited data. Such methods are important because they can be used to improve ASR systems for underrepresented groups and make it easier for them to experience the broad benefits of speech recognition technology, from virtual assistants to hands-free computing for the physically impaired.

2 Related Work

Data augmentation is the task of synthetically modifying data to increase the amount and diversity of the dataset. One method of conducting data augmentation for ASR is voice conversion (VC).

† Equal contribution

Voice conversion involves modifying the speaker characteristics of a given utterance while preserving linguistic information. Traditionally voice conversion methods have required the use of a large parallel training corpus, allowing training to be formulated as a regression problem. For example, Tanaka et al.'s work on sequence-to-sequence voice conversion using attention and context preservation mechanisms trains a converter between the source and target domain using parallel utterances in a supervised manner [2]. Similar works using full convolutional sequence-to-sequence networks have also relied on parallel training data [3]. The disadvantage of using a parallel corpus is that they are expensive to obtain and require time alignment preprocessing which can be difficult when there is a large acoustic gap between the source and target speech [4]. Obtaining a parallel corpus can be infeasible when working with low resource datasets.

Some works have explored training voice conversion systems on unpaired data. Sun et al. utilizes unpaired phonetic posteriorgrams (PPGs), to train a BiLSTM-RNN voice conversion model and Saito et al. proposed a VAE based framework for non-parallel conversion [5][6].

While VAEs and PPG frameworks have shown promise, CycleGAN based methods have widely been accepted as state-of-the-art approaches for non-parallel voice conversion [7]. Through adversarial training, CycleGAN learns a mapping from the source to target and the corresponding inverse mapping while being constrained by a cycle consistency loss term [8]. CycleGAN-VC, CycleGAN-VC2, and StarGAN-VCs have all demonstrated the ability to convert between source and target speakers [4][9][10]. Moreover, StarGAN-VCs are able perform many-to-many voice conversion by adding an encoding of the source and target speaker information at each layer [4]. However, these methods are constrained to conversions between Mel-frequency cepstrums (MFCCs) due to their limited ability to preserve time-frequency information between conversions [7].

To enable CycleGAN to capture time-frequency information, Kaneko et al. proposes CycleGAN-VC3 which incorporates a time-frequency adaptive normalization (TFAN) module. Using this module, they saw marked improvements in conversions between Mel-spectrograms. To reduce the number of parameters introduced by the TFAN module, Kaneko et al. instead proposed MaskCycleGAN-VC which masks out frames in the input Mel-spectrogram and trains the model on an auxiliary task to fill in the masked frames [7].

GAN-based data augmentation to improve downstream ASR systems is severely understudied. One of few studies in this domain utilizes CycleGAN to convert between child and adult speakers to address the paucity of publicly available child speech datasets [11]. Another study by Shahnawazuddin et al., demonstrate that augmenting with converted audio can improve ASR systems in a limited data scenario [11]. Gudepu et al. use a similar approach to convert between normal speech and whispered speech for data augmentation [12].

Our work has strong parallels to Shahnawazuddin et al. and Gudepu et al.'s publications. However, our work converts between generic American English and AAVE. Additionally, our work utilizes state-of-the-art MaskCycleGAN-VC to preserve time-frequency information.

3 Approach

To mitigate the performance gap of ASR systems trained on limited datasets of underrepresented speakers, our task in this project is to explore the effectiveness of GAN-based data augmentation to improve performance. In particular, we are focusing on improving the performance of an ASR system on the CoRAAL dataset [13] of African American speakers. In addition to being limited in size compared to other speech datasets, the CoRAAL dataset poses several key challenges that we attempt to mitigate in our approach.

AAVE speech in CoRAAL sounds distinctly different from larger datasets with generic American English speech. To address this, we use a GAN-based voice conversion model to convert speech from generic American English to AAVE, providing a proof-of-concept data augmentation method for low resource speech datasets. Additionally, there is some data imbalance between speakers (Appendix C Figure 6) in the CoRAAL dataset. To mitigate this imbalance, we augmented speakers with the least number of utterances to improve dataset diversity. To that end, we implemented and trained a CycleGAN-based voice conversion model to convert speech from 7 white speakers from VoC to match the style of the 7 African American speakers in CoRAAL with the least data.

3.1 CycleGAN-VC

3.1.1 Baseline

Voice conversion is the task of modifying the speaker characteristics of a given speech while preserving linguistic information. While current methods require expensive, large parallel training data, cycle-consistent generative adversarial networks (CycleGAN) are the state-of-the-art method in voice conversion that does not require a parallel corpus. CycleGAN is composed of four deep neural networks: two generators ($G_{A \to B}$ and $G_{B \to A}$) and two discriminators (D_A and D_B). The generator $G_{A \to B}$ is trained to convert input x from source domain A to target domain B. The outputs from the generator $G_{A \to B}(x)$ and real samples y from the target domain B are passed to the discriminator D_B. The discriminator D_B is trained to classify whether the input is real or fake. This adversarial objective allows the generator to learn to convert inputs from domain A to domain B in a realistic manner that fools the discriminator. More concretely, the adversarial loss function is defined as follows:

$$\mathcal{L}_{\text{adv}}(G, D) = \mathbb{E}_y[\log D(y)] + \mathbb{E}_x[\log(1 - D(G(x))] \tag{1}$$

The aforementioned adversarial loss is defined for both generators $G_{A \to B}$ and $G_{B \to A}$, which are jointly trained to convert between domain A and domain B. The inputs of our generator are 2D images known as Mel-spectrograms, which are generated during a preprocessing step from the raw wave file speech data (See section 4.1 for more details). In order to ensure that CycleGAN preserves linguistic information, a cycle consistency loss is added, which is defined as:

$$\mathcal{L}_{\text{cycle}}(G_{A \to B}, G_{B \to A}) = \mathbb{E}_x[\|G_{B \to A}(G_{A \to B}(x)) - x\|_1] + \mathbb{E}_y[\|G_{A \to B}(G_{B \to A}(y)) - y\|_1] \tag{2}$$

To ensure that the generator does not modify input speech if it is already in the target speaker's voice, an additional identity loss is added:

$$\mathcal{L}_{\text{identity}}(G_{A \to B}, G_{B \to A}) = \mathbb{E}_x[\|G_{B \to A}(x) - x\|_1] + \mathbb{E}_y[\|G_{A \to B}(y) - y\|_1] \tag{3}$$

3.1.2 Two step adversarial loss

While the baseline CycleGAN uses an L1 reconstruction loss between the real and reconstructed inputs, we follow the work of [9] to instead use a two step adversarial loss. The L1 reconstruction loss leads to a degradation in output quality due to statistical averaging, which is mitigated by using the two step adversarial loss. As such, we introduced two new discriminators $D_{A'}$ and $D_{B'}$ to add an adversarial loss between the real and reconstructed inputs.

Putting it all together, the combined training loss objective is defined as:

$$\begin{aligned}
\mathcal{L}_{\text{CycleGAN-VC}} = {} & \mathcal{L}_{\text{adv}}(G_{A \to B}, D_B) + \mathcal{L}_{\text{adv}}(G_{B \to A}, D_A) \\
& + \lambda_{\text{cycle}} \mathcal{L}_{\text{cycle}}(G_{A \to B}, G_{B \to A}) \\
& + \lambda_{\text{identity}} \mathcal{L}_{\text{identity}}(G_{A \to B}, G_{B \to A}) \\
& + \mathcal{L}_{\text{adv}}(G_{A \to B \to A}, D_{A'}) + \mathcal{L}_{\text{adv}}(G_{B \to A \to B}, D_{B'})
\end{aligned} \tag{4}$$

3.1.3 Model

The generator model, shown in figure 1, is composed of downsampling, residual and upsampling layers, following the model architecture proposed by [9]. We utilized a downsampling-upsampling architecture in order to lower the computational complexity. Furthermore, residual layers are an important component to combat the problem of vanishing gradients. In addition, our generator follows a 2-1-1D CNN architecture where the downsampling and upsampling blocks perform 2D convolutions, while the residual blocks perform 1D convolutions. Prior work [10] has shown that such an architecture would allow the model to effectively capture wide-range structures and features without degradation in performance.

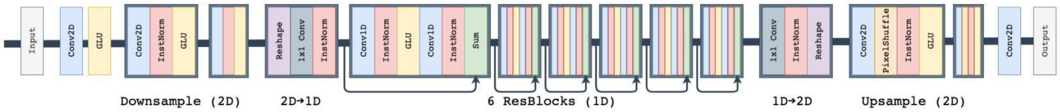

Figure 1: CycleGAN-VC Generator model

The discriminator model, shown in figure 2, is a fully convolutional PatchGAN architecture [14] that provides a real/fake prediction for each patch of the input Mel-spectrogram.

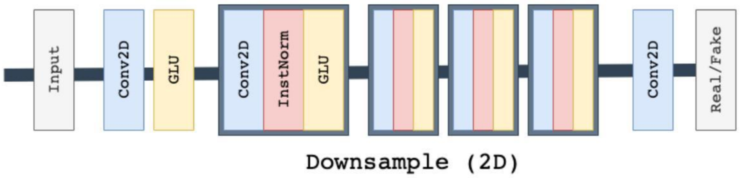

Figure 2: CycleGAN-VC Discriminator model

Inspired by existing implementations [15], we implemented the full CycleGAN-VC model along with its associated training, evaluation, logging and visualization scripts. Since most existing implementations of the CycleGAN-VC model had considerable bugs and limitations, we made significant contributions to rewrite and refactor extensive portions of prior implementations to fix correctness, improve clarity and adaptability.

3.1.4 Input Masking

To allow the CycleGAN model to grasp time-frequency structures of input Mel-spectrograms, we explored the use of an auxiliary self-supervised task as proposed in MaskCycleGAN-VC [7]. As shown in figure 3, a temporal mask is applied to the input Mel-spectrograms to encourage the generator to fill in missing frames using learned features from the surrounding frames. As such, this self-supervised task encourages the CycleGAN-VC model to learn the time-frequency structure of the input spectrograms without any additional learned parameters. Since the authors of [7] did not release their code, our repository has the first publicly available implementation of MaskCycleGAN-VC.

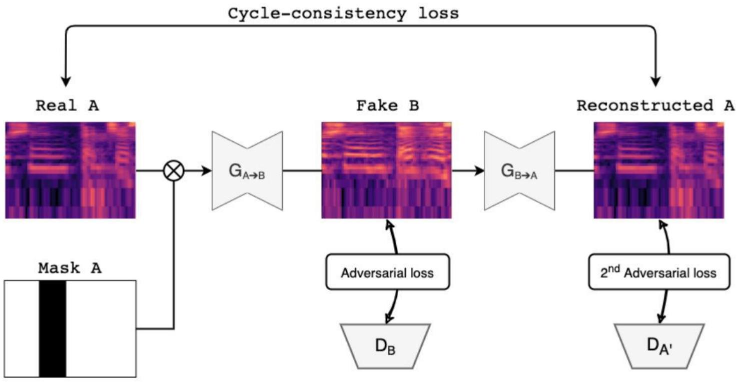

Figure 3: MaskCycleGAN-VC and its masked inputs

3.2 ASR

3.2.1 Model

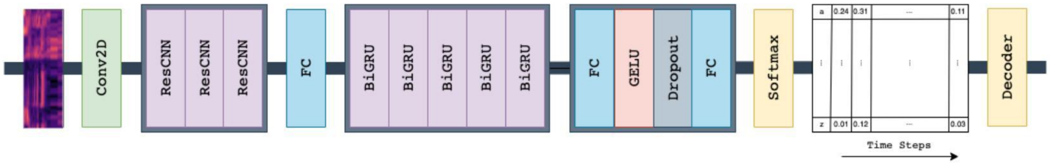

Figure 4: DeepSpeech2 ASR model architecture

We trained downstream ASR models to evaluate the utility of our converted samples for data augmentation. We utilized a DeepSpeech2 model architecture as shown in figure 4. The model consists of 3 residual CNN layers which utilize skip connections to stabilize gradient updates. The model is also comprised of 5 Bidirectional GRU layers which were selected over LSTM layers to reduce computational expenses. We significantly refactored AssemblyAI's implementation of DeepSpeech2 to incorporate it into our training pipeline [16].

3.2.2 Loss

We use a connectionist temporal classification (CTC) loss term to train our ASR model [17]. CTC loss computes the sum of probabilities over all possible alignments between the ASR model's prediction X and the ground truth label Y [18].

The softmax layer of the ASR model outputs probabilities for each time step which defines the distribution over output sequences $p_t(a|X)$ where t is the time step and a is a single alignment between X and Y. Valid output sequences which correspond to Y are obtained from a single alignment a by merging repeat characters, then removing ϵ tokens which are blank tokens [18]. Equation (5) describes the full CTC objective where for a training set $\mathcal{D}$, the model tries to minimize the negative log likelihood of each $(X\ Y)$ pair.

$$\mathcal{L}_{\text{CTC}} = \sum_{(X,Y)\in D} - \log p(Y|X) \tag{5}$$

Since the number of possible alignments represent such a large search space, the CTC conditional probability is estimated for each (X, Y) pair by using beam search to sum over the probabilities of the most probable valid alignments across time steps.

$$p(Y|X) = \sum_{A\in\mathcal{A}_{X,Y}} \prod_{t=1}^{T} p_t(a_t|X) \tag{6}$$

4 Experiments

4.1 Data

We use two distinct corpora of conversational speech. The first, called The Corpus of Regional African American Language (CoRAAL) [13], is a publicly available corpus of socio-linguistic interviews consisting of long-form, socio-linguistic interviews with black speakers from a variety of geographic areas.

The second corpus we use is called Voices of California (VoC) [19], which is similarly composed of socio-linguistic interviews, documenting variations in ways of speaking across different regions of California. We sampled 37 white speakers from the VoC dataset.

We preprocessed the audio from both datasets to splice out interviewers and created wave files of up to 20 seconds. For the ASR task, we normalized the transcripts according to Koenecke et al [1]. We split our CoRAAL data into train, validation and test sets by speaker, selecting 6 speakers of each gender ($\sim$15% of the total participants) at random for both the validation and test sets. Due to the variation in the quantity of data per speaker, we end up with a train set of 36.8 hours, a validation set of 6.8 hours and a test set of 7.3 hours. We used our entire VoC dataset, consisting of 37.0 hours of speech, for training.

During training, the input wave files were downsampled to 22.05 kHz. We extract 80-dimensional Mel-spectrograms with $L_{\text{window}} = 1024$, $L_{\text{hop}} = 256$.

4.2 Evaluation Method

Downstream ASR performance: We evaluate the performance of our downstream ASR model (DeepSpeech2) on the CoRAAL validation and test set using WER (Word Error Rate) and character-level Levenshtein distance (CER). We evaluate the effectiveness of our voice conversion data-augmentation by its impact on performance (WER, CER) of the ASR model on the CoRAAL dataset. Previous works on voice conversion have evaluated voice conversion using Mel-cepstral distortion and modulation-spectra distance, which requires paired utterances between speakers, which were unavailable between CoRAAL and VoC. Therefore we rely primarily on the performance of the downstream ASR task to evaluate the quality of our voice conversion.

Qualitative Measures: We used a pre-trained MelGAN vocoder [20] to convert Mel-spectrograms to wave forms, allowing us to validate the quality of the voice conversion.

4.3 Experimental Details

4.3.1 CycleGAN-VC

We experimented with three different versions of CycleGAN: CycleGAN-VC2, CycleGAN-VC3, and MaskCycleGAN-VC. For each version we trained three CycleGAN models to learn one-to-one mappings between three pairs of VoC and CoRAAL speakers. We decoded the converted Mel-spectrograms to waveform using a pre-trained MelGAN vocoder to evaluate the rendered audio. After determining the best model architecture, we trained 7 models to learn a mapping between 7 pairs of VoC and CoRAAL speakers to generate Mel-spectrograms for data augmentation. We carefully selected pairs of speakers to account for gender and age balance as well as clarity of audio.

For each speaker, we normalized the Mel-spectrograms using the mean and variance of the training set. During training, we randomly cropped 64 frames (~ 0.75 s) of the input Mel-spectrogram. We trained the model for 6172 epoch using an Adam optimizer, or until convergence, with a batch size of 1 on one Nvidia V100 Tensor Core GPU. The learning rates were set to 0.0002 for the generators and 0.0001 for the discriminators, with momentum terms $\beta_1 = 0.05, \beta_2 = 0.999$. Additionally, we used $\lambda_{\text{cycle}} = 10, \lambda_{\text{identity}} = 5$. After $10k$ steps, we decayed the learning rates linearly and removed the identity loss from the objective function so as not to over-constrain the generator. Our MaskCycleGAN-VC model randomly masked between 0 and 25 continuous frames.

4.3.2 ASR

We finetuned DeepSpeech2 ASR models which were pre-trained on 360 hours of "clean" speech from the LibriSpeech ASR corpus [21]. We utilized this finetuning approach as DeepSpeech2 struggled to learn from both CoRAAL and VoC datasets from scratch.

We trained each model for 100 epochs with a batch size of 10 on 2 Nvidia V100 Tensor Core GPUs. Each model took approximately 72 GPU hours to train. We utilized spectral augmentation [22] to increase the effective size of the dataset. We utilized an AdamW optimizer and a one cycle scheduler with a max learning rate of 0.0005 for training. We use greedy decoding to obtain our transcriptions.

4.4 Results

4.4.1 CycleGAN-VC

The outputs of the CycleGAN-VC model with two-step adversarial losses were qualitatively poor and choppy. Next we implemented MaskCycleGAN-VC by incorporating the auxiliary task of filling in masked frames. The outputs of MaskCycleGAN-VC were far superior to our prior models in terms of naturalness and similarity to the target speaker. As such, MaskCycleGAN-VC was selected as the architecture for the voice conversion task. We trained 7 MaskCycleGAN-VC models to learn a mapping between 7 pairs of VoC and CoRAAL speakers.

4.4.2 ASR

For our baselines, we trained our ASR model on the CoRAAL dataset (1), the unconverted VoC dataset (2) and the combination of the two datasets (3). Condition (2) acts as a proxy for traditional ASR models trained on generic American English.

For our first experiment with the converted data, we trained our ASR model with the CoRAAL dataset, the converted data from the 7 VoC speakers on which we trained our CycleGANs, and the unconverted VoC data from the remaining 29 speakers (4). Our model trained on this data performed worse than

Training Data	Val CER	Val WER	Val Loss
1. CoRAAL	0.2346	0.4981	1.382
2. Unconverted VoC	0.3698	0.6898	2.498
3. CoRAAL + unconverted VoC	**0.2191**	**0.4658**	1.18
4. CoRAAL + 7 converted VoC + 29 unconverted VoC	0.2209	0.4695	1.196
5. CoRAAL + 7 unconverted VoC	0.2274	0.4848	1.313
6. CoRAAL + 7 converted VoC	0.2335	0.4953	1.327
7. CoRAAL + unconverted VoC + 7 converted VoC	0.2226	0.4741	**1.089**

Table 1: DeepSpeech2 model performance on experimental conditions. We generated a one to one mapping between 7 VoC speakers and 7 CoRAAL speakers. We trained our CycleGAN models to convert the audio between these 7 VoC speakers and the 7 CoRAAL speakers. We then evaluated the ASR performance of our model trained on the listed combinations of our converted and unconverted data.

the model trained on CoRAAL and the unconverted VoC data (3). For our second experiment with the converted data, we trained our ASR model with the CoRAAL data and 7 unconverted VoC speakers as the control (5). As a comparison, we trained our ASR model with the CoRAAL data and audio from 7 VoC speakers converted using CycleGAN (6). This model saw a marked decrease across both metrics from the control (5). Finally, we train our ASR model on the CoRAAL data, all of the unconverted VoC data, and the converted VoC data from the selected 7 speakers (7). Despite training on the most data, this model performed worse than the baseline (3).

5 Analysis

5.1 Qualitative CycleGAN-VC results:

We converted spectrograms to waveforms using a pre-trained MelGAN vocoder to evaluate the quality of our voice conversion model. MaskCycleGAN-VC qualitatively produced the best voice conversion results across all testing speakers. In particular, it performed best when converting between two female speakers, and saw the worst results when converted cross-gender. This result is expected since the task of cross-gender voice conversion is inherently more difficult due to the larger difference in source and target domains.

Impact on ASR performance: As seen through our ablation studies in table 1, our method of CycleGAN voice conversion did not reliably improve the WER and CER of the ASR system on the validation set. When the number of utterances from VoC used for augmentation is fixed, ASR models trained on voice converted spectrograms were outperformed by their unconverted counterparts. This observation is seen across multiple parallel runs ((3)&(4) and (5)&(6)). This indicates that unconverted VoC spectrograms helped the model learn better than converted VoC spectrograms, and that improvements over the baseline when using converted spectrograms is a result of a larger dataset size. Experiment (7) further supports this observation, which demonstrates that the voice converted VoC spectrograms do not provide any additional information to improve ASR performance over the original unconverted spectrograms. This is a surprising result since the unconverted VoC spectrograms are from a very different distribution than the CoRAAL spectrograms that the model is being evaluated on. One possible explanation is that voice conversion reduces the diversity of the dataset by decreasing the number of total speakers in the dataset while generating more utterances for a fraction of CoRAAL speakers. Another possible explanation is that the converted audio quality is not as good as the unconverted audio. At times, we have noticed that the converted audio can be imperceptible and this may be leading to misalignment between audio samples and transcriptions, degrading model performance.

Transcriptions: Inspection of some transcriptions (see Appendix A) generated by the different models also provide some additional insights into the performance disparity between the models. The transcription generated by the model trained on the VoC dataset alone (3) only successfully transcribed the word "jump", exemplifying the vastly different distributions between the VoC and CoRAAL datasets. All other models were able to successfully transcribe the first five words "i don’t know how to" although the model trained on CoRAAL and unconverted VoC (5) was uniquely unable to distinguish "know" and "how" as separate words. We can also observe that the model trained on CoRAAL and converted VoC produces transcriptions that more closely resemble English words, despite having a higher CER and WER. Since the CER and WER metrics do not account for semantics of transcriptions, it is worth exploring an alternative metric that uses a language model for future work to better understand the impact of voice conversion augmentation.

Loss curves: As seen in Appendix B figure 5, our ASR models consistently began to overfit the training data after approximately 30 to 40 epochs. Despite an increasing validation loss, both CER and WER continued to decrease throughout training. To understand why this happens, we first note that CTC loss measures the probability across all valid alignments for a given target transcription. If our ASR model emphasizes a particular valid path, while penalizing other valid paths by making them less likely, it would result in an increase in the CTC loss. However, since the model is emphasizing a valid alignment that collapses to the correct transcription, the WER and CER metrics would decrease. This explains why there is a divergent behavior observed between the validation WER, CER and CTC loss.

Discrepancies in prior works: While CycleGAN is able to change the voice of speakers, it is unable to alter unique linguistic features of speech. Prior works which have found success conducting data augmentation through voice conversion, such as conversion between normal speech and whispered speech, have addressed data scarcity by converting speech between domains that differ more superficially. However, the CoRAAL and VoC datasets differ significantly in both voice and content due to fundamental linguistic differences between AAVE and generic American English. This suggests the need for an intermediate step to translate between generic American English and AAVE before conducting voice conversion. Given that no such parallel corpus exists, we would have to rely on unsupervised machine translation for this task.

6 Conclusion

6.1 Main Findings

Our main finding was that using Voice Conversion as a data augmentation technique did not improve the performance of our ASR model due to the large linguistic and contextual discrepancy between the CoRAAL and VoC datasets. One of the main highlights of our project was developing the first publicly available implementation of CycleGAN-VC3 and MaskCycleGAN-VC which holds true to the original paper. One of the limitations of our work was that we only conducted voice conversion among a subset of the CoRAAL and VoC dataset. If time and compute permits, it would be highly rewarding to learn mappings between each speaker in both datasets to investigate what factors make converted audio useful for data augmentation to develop an optimal curriculum. While this project was challenging, it was especially rewarding to conduct this line of research which has the ultimate goal of ensuring that marginalized voices are heard.

6.2 Future Work

For future work, we could experiment with training CycleGAN to convert between CoRAAL speakers. This would enable internal augmentation and would help rebalance the dataset. Converting between CoRAAL speakers would ensure that the converted utterances maintain linguistic and semantic contexts characteristic of the CoRAAL dataset. Additionally, we could train a many-to-many voice conversion model as done by Kameoke et al., as our single voice conversion model. While this model may take longer to train, it would considerably streamline our data augmentation pipeline which currently relies on training one model for every pair of speakers.

A method that could be used to improve the transcriptions from our model is to use beam search decoding with a language model. The language model could be trained on the CoRAAL dataset, or could also be pre-trained on a larger corpus. We expect that incorporating a language model would immediately improve WER and CER for each of our models.

References

[1] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[2] Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, and Nobukatsu Hojo. Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms, 2018.

[3] Hirokazu Kameoka, Kou Tanaka, Damian Kwasny, Takuhiro Kaneko, and Nobukatsu Hojo. Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion, 2020.

[4] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–273. IEEE, 2018.

[5] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016.

[6] Yuki Saito, Yusuke Ijima, Kyosuke Nishida, and Shinnosuke Takamichi. Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5274–5278. IEEE, 2018.

[7] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Maskcyclegan-vc: Learning non-parallel voice conversion with filling in frames, 2021.

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.

[9] Takuhiro Kaneko and Hirokazu Kameoka. Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2100–2104. IEEE, 2018.

[10] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6820–6824. IEEE, 2019.

[11] S Shahnawazuddin, Nagaraj Adiga, Kunal Kumar, Aayushi Poddar, and Waquar Ahmad. Voice conversion based data augmentation to improve children's speech recognition in limited data scenario. *Proc. Interspeech 2020*, pages 4382–4386, 2020.

[12] Prithvi RR Gudepu, Gowtham P Vadisetti, Abhishek Niranjan, Kinnera Saranu, Raghava Sarma, M Ali Basha Shaik, and Periyasamy Paramasivam. Whisper augmented end-to-end/hybrid speech recognition system-cyclegan approach. *Proc. Interspeech 2020*, pages 2302–2306, 2020.

[13] Tyler Kendall and Charlie Farrington. The corpus of regional african american language. 2020.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[15] Kun Ma. Cyclegan-vc3. `https://github.com/jackaduma/CycleGAN-VC3`, 2020.

[16] Building an end-to-end speech recognition model in pytorch.

[17] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[18] Awni Hannun. Sequence modeling with ctc. *Distill*, 2017. https://distill.pub/2017/ctc.

[19] Stanford Linguistics. Voices of california.

[20] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis, 2019.

[21] Vassil Panayotov, 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)*, pages 5206–5210. IEEE, 2015.

[22] Daniel S. Park, 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. *Interspeech 2019*, Sep 2019.

A Sample Generated Transcriptions from ASR Models

Data	Transcription
Ground Truth	i don’t know how to double dutch i know how to hopscotch and jump regular rope
1. CoRAAL	i don’t know how to snumle dudche i lno how a hope scotch and jump breula road
2. Unconverted VoC	atta glot as a wit this a’m not o has spatch an jump regu lerolte
3. CoRAAL + unconverted VoC	i don’t knowhow to nobleh thosh i notw to hopscotch an jump bregula road
4. CoRAAL + 7 converted VoC + 29 unconverted VoC	i don’t know how to numble o dos i now how to hape scotch and jump bregularo
5. CoRAAL + 7 unconverted VoC	i don’t know how to nem be doudch i no how to hop scotch and jump breular road
6. CoRAAL + 7 converted VoC	i don’t know how to numble doch i’m low how to hope schotch and jump regular role
7. CoRAAL + unconverted VoC + 7 converted VoC	i don’t know how to no wi dos i now how to hap schotch and jump regula roop

Table 2: DeepSpeech model transcriptions on CoRAAL and Combined data

B Training Loss

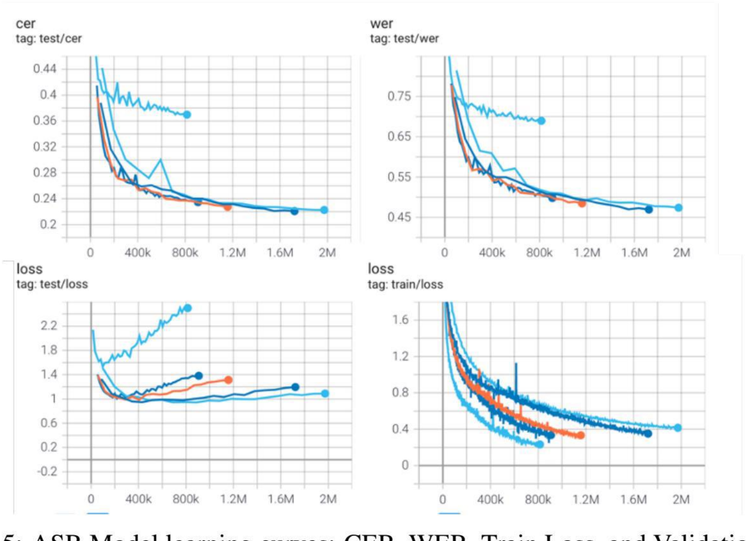

Figure 5: ASR Model learning curves: CER, WER, Train Loss, and Validation Loss

C Data Distribution

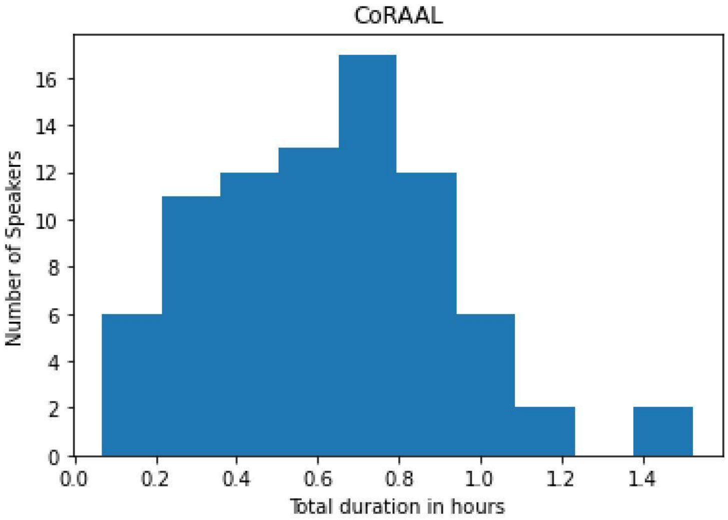

Figure 6: Distribution of total duration for speakers from the CoRAAL dataset

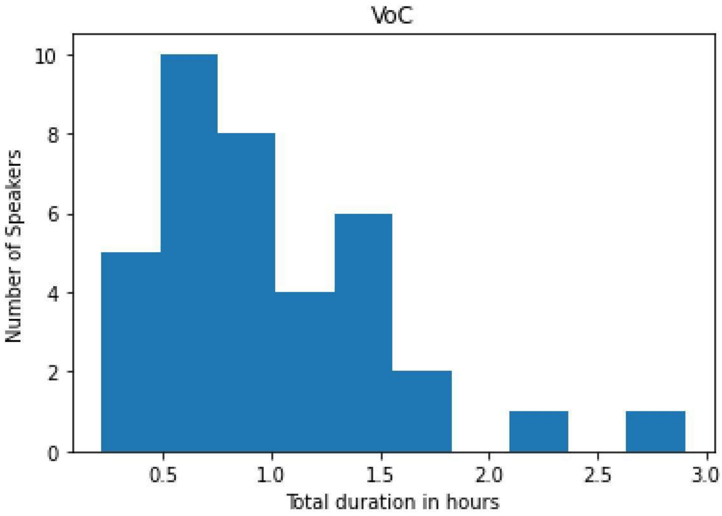

Figure 7: Distribution of total duration for speakers from the VoC dataset