Mitigating bias against non-native accents

Zhang, Yuanyuan; Zhang, Yixuan; Halpern, Bence Mark; Patel, Tanvina; Scharenborg, Odette

DOI
10.21437/Interspeech.2022-836

Publication date
2022

Document Version
Final published version

Published in
Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH

Interspeech 2022
18-22 September 2022, Incheon, Korea

Mitigating bias against non-native accents

Yuanyuan Zhang[1], *Yixuan Zhang*[1], *Bence Mark Halpern*[1, 2, 3], *Tanvina Patel*[1], *Odette Scharenborg*[1]

[1]Multimedia Computing Group, Delft University of Technology, Delft, the Netherlands
[2]Netherlands Cancer Institute, Amsterdam, the Netherlands
[3]ACLC, University of Amsterdam, Amsterdam, the Netherlands

(y.zhang-88, y.zhang-96)@student.tudelft.nl, (T.B.Patel, O.E.scharenborg)@tudelft.nl, B.M.Halpern@uva.nl

Abstract

Automatic speech recognition (ASR) systems have seen substantial improvements in the past decade; however, not for all speaker groups. Recent research shows that bias exists against different types of speech, including non-native accents, in state-of-the-art (SOTA) ASR systems. To attain inclusive speech recognition, i.e., ASR for everyone irrespective of how one speaks or the accent one has, bias mitigation is necessary. Here we focus on bias mitigation against non-native accents using two different approaches: data augmentation and by using more effective training methods. We used an autoencoder-based cross-lingual voice conversion (VC) model to increase the amount of non-native accented speech training data in addition to data augmentation through speed perturbation. Moreover, we investigate two training methods, i.e., fine-tuning and domain adversarial training (DAT), to see whether they can use the limited non-native accented speech data more effectively than a standard training approach. Experimental results show that VC-based data augmentation successfully mitigates the bias against non-native accents for the SOTA end-to-end (E2E) Dutch ASR system. Combining VC and speed perturbed data gave the lowest word error rate (WER) and the smallest bias against non-native accents. Fine-tuning and DAT reduced the bias against non-native accents but at the cost of native performance.

Index Terms: speech recognition, bias mitigation, voice conversion, data augmentation, domain adversarial training

1. Introduction

Automatic speech recognition (ASR) has improved a lot since the introduction of deep learning techniques [1–8]. Currently, the state-of-the-art (SOTA) ASR systems work extremely well for speakers whose speech patterns match its training data: typically, these are adult highly-educated first-language speakers of a standardized dialect, with little or no speech disability (referred to as norm speakers). Anecdotal and recent empirical evidence, however, have shown that for many groups of people ASR works less well [9, 10], even when the ASR systems are trained on the speech of that speaker group [11]. In other words, SOTA ASR systems are biased against speakers whose speech deviates from norm speakers. For instance, recent experimental results in [9] showed a bias against non-native accents in both read speech and human-machine interaction (HMI) dialogue speech for children and adult speakers. In this paper, we focus on mitigating bias against non-native accented speech in a SOTA end-to-end (E2E) ASR system, where we define bias as the performance gap between accented and non-accented speech. Specifically, we will work on the Dutch ASR system for both read and conversational type of speech.

There is very limited research focusing on bias mitigation against non-native accented speech. Instead, most existing research focuses on improving non-native accented speech recognition, i.e., lowering the word error rate (WER) or character error rate (CER) [12–18]. However, we aim to build inclusive ASR, i.e., ASR for everyone irrespective of how one speaks, so our aim should not only be to reduce WER but also to reduce the performance gap between non-accented and accented speech recognition. For example, in [13], the WER on non-native accented speech improved while at the same time the bias against the non-native accented speech increased, because the native WER improved more. Non-native accented speech recognition is typically a low-resource problem, i.e., very little non-native accented training data is available. In this paper, we focus on bias mitigation against non-native accents using two different approaches: data augmentation to increase the amount of training data and by using more effective training methods.

Common data augmentation techniques include SpecAugment [19], speed perturbation [20], and reverberation [21], which modify the speech or spectrogram directly, but ignore some unique aspects of human speech [22]. To address this, voice conversion (VC) can be used to generate more speech data with different voices. The VC-generated speech data has the potential to improve ASR performance, e.g., in [22], VC was successfully used to augment the natural speech for four kinds of languages in very low-resource settings. In [23], VC was used to augment child speech data which improved child speech recognition performance. Thus, VC has the potential to augment non-native accented speech data, and improve non-native accented speech recognition. Furthermore, in order to make ASR systems more robust to non-native accents, various training methods have been proposed. For instance, domain adversarial training (DAT) [18] improved the performance of accented speech recognition for both E2E [17] and hybrid ASR systems [18]. Combining DAT with transfer learning further improved the performance [14].

In this paper, we used an autoencoder-based cross-lingual VC model to increase the amount of non-native accented speech training data in addition to using speed perturbation for data augmentation. Moreover, we applied two training methods, i.e., DAT and fine-tuning, to investigate whether they can be used to train the limited non-native accented speech data more effectively than a standard training approach.

2. Methodology

We trained a baseline SOTA E2E model for Dutch (see Section 2.3.1) and tested it on native and non-native accented Dutch (see Section 2.1). In order to mitigate bias against non-native accented speech, we investigated different combinations of cross-lingual VC-based non-native accented speech augmentation and

10.21437/Interspeech.2022-836

speed perturbation data augmentation (Section 2.2) and DAT and fine-tuning training (Section 2.3) techniques.

2.1. Datasets

2.1.1. The spoken Dutch corpus (CGN)

Corpus Gesproken Nederlands-CGN [24] is a Dutch corpus containing native speech data spoken by speakers from the Netherlands and Flanders. We only use the data recorded in the Netherlands to train our E2E ASR systems. In this study, we followed the training and test sets used in [9]. The CGN training data was processed into small chunks and cleaned using the Kaldi toolkit yielding 380.12 hours of standard Dutch speech data, denoted by $\mathbf{C_{train}}$. The CGN data consists of two test sets, i.e., broadcast news)BN) and conversational telephone speech (CTS) as follows:

- $\mathbf{C_{BN}}$: broadcast news (0.4 hours)
- $\mathbf{C_{CTS}}$: conversational telephone speech (1.8 hours)

2.1.2. Jasmin-CGN corpus

Jasmin-CGN corpus [25] is an extension of the CGN corpus consisting of read speech and HMI speech spoken by native speakers (children, teenagers, older adults) and non-native speakers (children, adults). Furthermore, the non-native speakers come from 37 different countries such as Afghanistan, Andorra, Egypt and Spain, etc. We divided the Jasmin-CGN corpus into a training set and 4 test sets. The 4 test sets are (In each test set, half of the speakers are male and half are female):

- $\mathbf{R_N}$: native read speech (1.45 hours),
- $\mathbf{R_{NN}}$: non-native read speech speakers (1.63 hours),
- $\mathbf{H_N}$: native HMI speech (0.68 hours),
- $\mathbf{H_{NN}}$: non-native accented HMI speech (0.36 hours).

The training set contains 36.12 hours speech data, denoted by $\mathbf{J_{train}}$, including 14.1 hours non-native accented speech (10.42 hours read data and 3.69 hours HMI data) and 22.02 hours native speech (16.31 hours read data and 5.70 hours HMI data).

2.1.3. VCTK corpus

The VCTK corpus [26] is an English multi-speaker corpus, consisting of speech from 109 English speakers (62 male speakers and 47 female speakers) with various accents. We will use the VCTK corpus for our VC experiments.

2.2. Data Augmentation

Data augmentation is only applied to the original non-native accented speech data in $\mathbf{J_{train}}$ because we have relatively large amounts of native speech data.

2.2.1. Cross-lingual VC-based non-native accents

For the VC experiments, we use a SOTA non-parallel VC model: AGAIN-VC [27][1]. It is an autoencoder-based VC model which disentangles the speaker and content information of the input speech data. In contrast to the original paper [27], we use AGAIN-VC cross-lingually. The reason is that we need to ensure that the generated VC speech contains non-native accented speech characteristics. Converting Dutch native speech as the source while using the Jasmin non-native accented speech or the VCTK English speakers as the target will not ensure this, because voice conversion increases the speaker variation but does not change the accent of the source speech data. Therefore, we convert the voices of the non-native Dutch speakers from the Jasmin corpus to those of the English speakers in the VCTK dataset, which will ensure that the converted speech is non-native accented. To that end, in contrast to the original paper [27], we train the VC model with more training data (more speakers) and inference with two languages (English, Dutch) speakers rather than one (English). In order to ensure that the intelligibility and quality of the generated non-native accented speech data are good, we use the cosine speaker similarity method [28] to find source speakers (non-native Dutch speakers, keeping the content information of the speech) and target speakers (English speakers in VCTK, keeping the speaker information of the speech) which are relatively similar.

For data processing, we follow the experimental setup as in [27]. The waveforms were first downsampled to a sampling frequency of 22050Hz, after which silence at the start and end of each audio clip were removed. Next, the mel-spectrogram features of the audio clips were extracted with 1024 window size, 256 hop length, and 80 mel bins. We use the full VCTK dataset (109 English speakers) and all non-native accented speech data in J_{train} (86 Dutch speakers) to train the AGAIN-VC model. The mel-spectrogram of each speech data was input to the encoder, where the mel-spectrogram was disentangled into two parts: speaker information (voice) and the linguistic content information. After encoding, both the speaker information and the content information were passed to the decoder, where the new mel-spectrogram was reconstructed by combining the speaker information and the content information. The AGAIN-VC only uses the mean absolute error as the self-reconstruction loss [27]. Following [27], the batch size was set to 32, and the number of training steps was 100k.

The output of the AGAIN-VC model is the mel-spectrogram of the converted speech data. As in [27], the output mel-spectrogram of the VC model is re-synthesised using a pre-trained Mel-GAN vocoder and we downsampled the generated speech data to 16000Hz for the subsequent ASR experiments. As a result, we generated 4.25 hours of non-native HMI data denoted by $\mathbf{vc_{hmi}}$ and 19.13 hours of non-native read data denoted by $\mathbf{vc_{rd}}$, using 192 pairs of source-target speakers. $\mathbf{vc_{all}}$ denotes the VC training set that combines $\mathbf{vc_{hmi}}$ and $\mathbf{vc_{rd}}$.

2.2.2. Speed perturbation for non-native accents

We used the standard speech perturbation [20] data augmentation method: the speed command of *sox* is used to do two-fold speed perturbation data augmentation (with 0.9 and 1.1 perturbation factors) for both the non-native accented read speech data and HMI data, respectively denoted by $\mathbf{sp_{rd}}$ and $\mathbf{sp_{hmi}}$. $\mathbf{sp_{all}}$ indicates the combination of $\mathbf{sp_{rd}}$ and $\mathbf{sp_{hmi}}$.

2.3. ASR Training Strategies

2.3.1. Baseline model

The SOTA ASR model is a transformer-based sequence-to-sequence (seq2seq) ASR model (referred to as T in Table 1). It is trained using the transformer recipe of LibriSpeech in the SpeechBrain toolkit [29]. Figure 1 shows the network configuration of our baseline model (without the domain classifier). The encoder consists of 3 CNN blocks and a 12-layer transformer. The decoder is composed of a 6-layer transformer with 8 heads. After inputting the training speech data, the 80-

[1] https://github.com/KimythAnly/AGAIN-VC

dimensional mel-spectrogram was calculated as the input feature denoted by X. The output of the encoder consists of high level features. The CTC loss and the transformer decoder both use the same high level features to achieve joint CTC-Attention decoding. All experiments share the same network configuration as the baseline.

The baseline model is trained using the Dutch data from $\mathbf{C_{train}}$. Moreover, to investigate the effect of non-native accented speech data augmentation, the training data $\mathbf{C_{train}}$ was augmented with: a) $\mathbf{J_{train}}$; b) $\mathbf{J_{train}}$ and $\mathbf{sp_{rd}}$; c) $\mathbf{J_{train}}$ and $\mathbf{sp_{hmi}}$; d) b+c; e) $\mathbf{J_{train}}$ and $\mathbf{vc_{rd}}$; f) $\mathbf{J_{train}}$ and $\mathbf{vc_{hmi}}$; g) e+f; h) d+g.

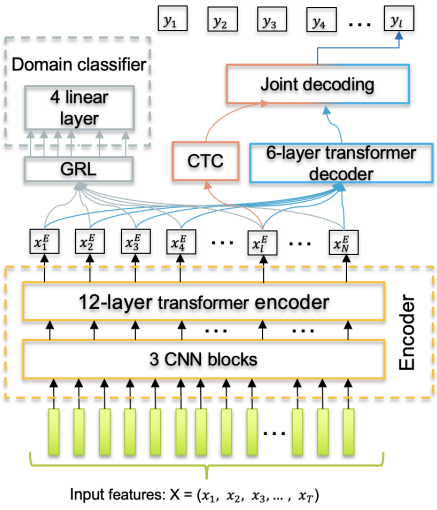

Figure 1: *Transformer-based ASR system with(out) DAT. Baseline ASR: without the domain classifier; ASR combined with the DAT training strategy: with the domain classifier.*

2.3.2. Training strategies

We investigate the effect of fine-tuning and DAT on bias mitigation of non-native accented speech.

Fine-tuning is a transfer learning approach where a neural network is retrained starting from a pre-trained model. For the fine-tuning experiments, we take the baseline E2E model trained on the CGN training data $\mathbf{C_{train}}$ as the pre-trained model. We fine-tuned the ASR model for 5 epochs using the a) speech data from $\mathbf{J_{train}}$; b) speech data from $\mathbf{J_{train}}$, $\mathbf{vc_{all}}$; c) speech data from $\mathbf{J_{train}}$, $\mathbf{vc_{all}}$ and $\mathbf{sp_{all}}$.

DAT is a kind of domain adaption method, aiming to extract domain-invariant features. To investigate the effect of DAT on bias mitigation, a domain classifier was added to our baseline ASR model, see top-left of Figure 1. The domain classifier is a binary classifier composed of 4 linear layers which share the same features with the decoder and is used to classify whether the input data is spoken by a native speaker or a non-native speaker. With the help of a gradient reversal layer (GRL), we can change the features extracted by the encoder, making the features accent-invariant.

The loss of the ASR model and the loss of the domain classifier are denoted by $Loss_{ASR}$ and $Loss_{domain}$, respectively. With DAT, the total loss is: $Loss_{DAT} = Loss_{ASR} + \lambda Loss_{domain}$, where $\lambda \in \mathbb{R}$ is a hyper-parameter, which controls the degree of influence of the domain classifier on the whole model. In our DAT experiments, we set λ to 0.01, the same value as [17]. In addition, we use supervised-DAT, which means that except for training the ASR model with native speech data, we also use the non-native accented speech data to train the ASR model because we have the transcriptions of the non-native accented speech data and the experimental results in [18] suggest a benefit from supervised-DAT. We trained three DAT models: the standard training set $\mathbf{C_{train}}$ augmented with a) speech data from $\mathbf{J_{train}}$; b) speech data from $\mathbf{J_{train}}$, $\mathbf{vc_{all}}$; c) speech data from $\mathbf{J_{train}}$, $\mathbf{vc_{all}}$ and $\mathbf{sp_{all}}$.

2.3.3. Evaluation

The performance is reported in terms of WER for the native and non-native accented speech and for read speech and human-machine interaction speech separately. The native speech is evaluated on the CGN BN and CTS test sets ($\mathbf{C_{BN}}$, $\mathbf{C_{CTS}}$) and the Jasmin test sets ($\mathbf{R_N}$, $\mathbf{H_N}$). The non-native accented speech is evaluated on the Jasmin test sets ($\mathbf{R_{NN}}$, $\mathbf{H_{NN}}$). Bias against non-native accents is conceptualized as the gap between the WER performance on the native speech and the non-native accented speech, and is calculated as the difference between the WER on the non-native accented speech and the WER on the native speech of the Jasmin corpus. For read data, the bias is denoted as $\mathbf{B_R}$; for HMI data, the bias is denoted as $\mathbf{B_H}$.

3. Results and Discussion

Table 1 shows the results in terms of WER and bias for the Baseline model and the data augmentation experiments (T), and fine-tuning (F-T) and DAT experiments on different data sets. The baseline model shows a bias against non-native accented read speech of 28.83% and a bias of 29.49% for HMI speech.

3.1. Data Augmentation Results

Table 1 shows the baseline results and the 8 data augmentation results (indicated by a-h) in Section 2.3.1. Adding $\mathbf{J_{train}}$ to the training data reduced WER and bias against non-native accented speech on the Jasmin tests substantially for both read speech (by 13.04%) and HMI speech (by 13.11%) compared to the baseline. This reduction in WER was not observed for the native speech in the CGN test sets. Adding speed perturbed or VC non-native accented speech further improved recognition performance on the Jasmin tests and reduced the bias, with both techniques giving similar performance. Adding read speech data leads to the biggest bias reduction for read speech; while adding HMI speech leads to the biggest bias reduction for HMI speech. The best results in terms of both WER and bias were obtained when generated read and HMI speech were both added. For the native speech in the CGN test sets, performance did not change much when data augmentation of non-native accented speech was applied, although also here a small improvement could be observed when all generated data was added. Compared to the baseline, our best model reduced WER overall (final column) with 14.48% and, crucially, reduced bias by 15.24% for read data and by 16.87% for HMI data.

CGN and Jasmin are two different databases, and training and test set mismatches might lead to reduced performance. Moreover, we wanted to check the quality of the VC-generated non-native accented speech. To that end, we also trained our system using only the Jasmin data, and with adding $\mathbf{vc_{all}}$, $\mathbf{sp_{all}}$ and both $\mathbf{vc_{all}}$ and $\mathbf{sp_{all}}$. Table 2, shows the results: 1) the reduction in training data when removing the CGN training data reduced performance. Moreover, performance improved when VC-speech was added, although speed perturbation led to bet-

Table 1: *Experimental results (WER and bias) of **T**: standard training, with(out) data augmentation; **F-T**: fine-tuning of the baseline model trained on $\mathbf{C_{train}}$ on other data; **DAT**: DAT experiments, on the native speech from CGN and for the native and non-native speaker groups from Jasmin, and for read speech and spontaneous/HMI speech, separately. $\mathbf{R/H_N}$: Read/HMI speech, native speakers in Jasmin; $\mathbf{R/H_{NN}}$: Read/HMI speech, non-native speakers in Jasmin; $\mathbf{B_{R/H}}$: **B**ias against non-native accents for read/HMI speech; $\mathbf{A_{Jas}}$: Averaged WER across native and non-native speakers in Jasmin.*

Details			CGN (% WER)		Jasmin (% WER and Bias)						
Model	**Training data**	**hours**	$\mathbf{C_{BN}}$	$\mathbf{C_{CTS}}$	$\mathbf{R_N}$	$\mathbf{R_{NN}}$	$\mathbf{B_R}$	$\mathbf{H_N}$	$\mathbf{H_{NN}}$	$\mathbf{B_H}$	$\mathbf{A_{Jas}}$
Baseline T	$\mathbf{C_{train}}$	380.12	9.64	37.99	24.9	53.73	28.83	30.77	60.26	29.49	38.64
T	$\mathbf{C_{train}, J_{train}}$	416.24	9.75	37.01	5.37	21.16	15.79	20.12	36.50	16.38	16.13
T	$-, \mathbf{sp_{rd}}$	437.08	10.08	37.54	4.90	18.78	13.88	20.77	35.89	15.12	15.24
T	$-, \mathbf{sp_{hmi}}$	423.62	9.61	37.25	5.28	20.70	15.42	20.45	34.20	13.75	15.75
T	$-, \mathbf{sp_{all}}$	444.46	9.59	36.72	4.98	18.75	13.77	20.22	33.16	12.94	14.87
T	$-, \mathbf{vc_{rd}}$	435.37	10.84	38.99	5.31	19.49	14.18	21.09	37.90	16.81	15.91
T	$-, \mathbf{vc_{hmi}}$	420.49	9.50	37.45	5.55	20.48	14.93	20.53	35.43	14.90	15.94
T	$-, \mathbf{vc_{all}}$	439.62	9.46	37.02	4.88	18.83	13.95	20.80	34.32	13.52	15.08
T	$-, \mathbf{vc_{all}}, \mathbf{sp_{all}}$	467.84	**9.27**	36.53	4.79	**18.38**	13.59	**19.62**	**32.24**	12.62	**14.48**
F-T	$\mathbf{C_{train}, J_{train}}$		30.93	48.51	5.00	20.42	15.42	21.27	35.26	13.99	15.80
F-T	$-, \mathbf{vc_{all}}$		39.18	53.71	**4.75**	19.73	14.98	22.80	34.78	11.98	15.69
F-T	$-, \mathbf{vc_{all}}, \mathbf{sp_{all}}$		43.76	55.74	4.78	19.25	14.47	23.90	34.03	**10.13**	15.65
DAT	$\mathbf{C_{train}, J_{train}}$		11.00	40.89	6.12	21.81	15.69	22.25	38.77	16.52	17.25
DAT	$-, \mathbf{vc_{all}}$		10.95	39.94	5.40	19.72	14.32	22.19	34.68	12.49	15.84
DAT	$-, \mathbf{vc_{all}}, \mathbf{sp_{all}}$		10.23	**32.87**	4.86	18.40	**13.54**	20.24	32.87	12.63	14.69

ter results. The best results were obtained when both VC-based and speed perturbed speech were added.

Table 2: *Experimental results (WER) when only using Jasmin data for the standard training method*

Training data	$\mathbf{A_{Jas}}$
$\mathbf{J_{train}}$	56.71
$\mathbf{J_{train}, vc_{all}}$	46.43
$\mathbf{J_{train}, sp_{all}}$	35.90
$\mathbf{J_{train}, vc_{all}, sp_{all}}$	31.58

3.2. Experimental Results of Fine-tuning and DAT

The **Fine-tune** results are shown in Table 1 (rows with F-T). Compared with the baseline, both WER and bias are reduced when fine-tuning using the Jasmin data is applied for native and non-native accented speech in the Jasmin corpus. However, WER increases drastically for the native speakers in CGN. Performance further improves for the Jasmin test sets (and decreases for CGN) when the fine-tuning data is increased with VC speech and both the VC speech and the speed perturbed speech, the latter yielding the best results, with a bias reduction of 14.36% for read speech, and 19.36% for HMI speech - but note that this latter result comes at the cost of an increase in the WER for the native speakers in Jasmin.

The **DAT** results are quite similar to those for fine-tuning with, except that the decrease in WER for the native speech in CGN is less big. Again, the best result is obtained when adding the VC and speed perturbed speech, leading to a bias reduction of 15.29% for read data, and 16.86% for HMI data.

Comparing all results showed that the best non-native accent results and overall WER were obtained when using a standard training approach with both VC and speed perturbed data added. The smallest bias for read speech was found for DAT combined with both data augmentation approaches. The smallest bias for HMI speech was observed for fine-tuning with both data augmentation (but at the cost of native performance).

Although DAT improved performance, the improvement is smaller than in [18] and [17]. Their amount of accented speech data was however substantially larger than ours. Moreover, we used their value of the hyper-parameter λ. Finding the optimal value for our database is to be explored in the future.

4. Conclusions and Future Work

In this paper, we aim to reduce bias against non-native accents using augmentation techniques and by exploring alternate training methods. The results showed that both cross-lingual voice conversion based data augmentation and speed perturbation lead to the improvement of non-native accented speech recognition performance and reductions in bias against non-native accents for all training methods. A combination of VC and speed perturbed data gave the lowest WER and smallest bias. Comparison of the standard training approach, fine-tuning, and domain adversarial training showed that that the standard training approach gave the best results. The best model was trained with the combination of VC and speed perturbed speech with standard training method, and reduced the non-native bias for read data from 28.83% to 13.59%, and for HMI data from 29.49% to 12.62% simultaneously.

Future work will focus on further exploring the cross-lingual VC and the speed perturbation. For both techniques we used the same amount of data in all experiments, while different amounts of additional VC-augmented data and speed perturbed data are worth exploring. Moreover, it is worth investigating whether different E2E and hybrid architectures will lead to further reductions in bias against non-native accented speech.

5. References

[1] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 244–250, 2021.

[2] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.

[3] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli, "Self-training and pre-training are complementary for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3030–3034.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[5] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, "Speechstew: Simply mix all available speech recognition data to train one large neural network," *arXiv preprint arXiv:2104.02133*, 2021.

[6] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition." in *INTERSPEECH*, 2020, pp. 2817–2812.

[7] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," *arXiv preprint arXiv:2005.09267*, 2020.

[8] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition." in *INTERSPEECH*, 2020, pp. 5036–5040.

[9] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, 2021.

[10] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[11] T. Kendall and C. Farrington, "The corpus of regional african american language," *Version*, vol. 6, p. 1, 2018.

[12] T. Han, H. Huang, Z. Yang, and W. Han, "Supervised contrastive learning for accented speech recognition," *arXiv preprint arXiv:2107.00921*, 2021.

[13] N. Das, S. Bodapati, M. Sunkara, S. Srinivasan, and D. H. Chau, "Best of both worlds: Robust accented speech recognition with adversarial transfer learning," in *INTERSPEECH*, 2021, pp. 1314–1318.

[14] Y.-C. Chen, Z. Yang, C.-F. Yeh, M. Jain, and M. L. Seltzer, "Aipnet: Generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6979–6983.

[15] H. Hu, X. Yang, Z. Raeesy, J. Guo, G. Keskin, H. Arsikere, A. Rastrow, A. Stolcke, and R. Maas, "Redat: Accent-invariant representation for end-to-end asr by domain adversarial training with relabeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6408–6412.

[16] T. Viglino, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition." in *INTERSPEECH*, 2019, pp. 2140–2144.

[17] H.-J. Na and J.-S. Park, "Accented speech recognition based on end-to-end domain adversarial training of neural networks," *Applied Sciences*, vol. 11, no. 18, p. 8412, 2021.

[18] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4854–4858.

[19] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019, pp. 2613–2617.

[20] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[21] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[22] M. Baas and H. Kamper, "Voice conversion can improve asr in very low-resource settings," *arXiv preprint arXiv:2111.02674*, 2021.

[23] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice conversion based data augmentation to improve children's speech recognition in limited data scenario." in *INTERSPEECH*, 2020, pp. 4382–4386.

[24] N. Oostdijk *et al.*, "The spoken dutch corpus. overview and first evaluation." in *LREC*. Athens, Greece, 2000, pp. 887–894.

[25] C. Cucchiarini, H. V. Hamme, O. v. Herwijnen, and F. Smits, "Jasmin-cgn: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality," 2006.

[26] M. K. Yamagishi Junichi, Veaux Christophe, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[27] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H.-y. Lee, "Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization," in *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5954–5958.

[28] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[29] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.