Exploring Data Augmentation in Bias Mitigation Against Non-Native-Accented Speech

Zhang, YuanYuan ; Herygers, Aaricia ; Patel, Tanvina; Yue, Zhengjun; Scharenborg, Odette

DOI
10.1109/ASRU57964.2023.10389756

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)

Citation (APA)
Zhang, Y., Herygers, A., Patel, T., Yue, Z., & Scharenborg, O. (2023). Exploring Data Augmentation in Bias Mitigation Against Non-Native-Accented Speech. In *Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* IEEE. https://doi.org/10.1109/ASRU57964.2023.10389756

2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 979-8-3503-0689-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASRU57964.2023.10389756

EXPLORING DATA AUGMENTATION IN BIAS MITIGATION AGAINST NON-NATIVE-ACCENTED SPEECH

Yuanyuan Zhang[1], *Aaricia Herygers*, *Tanvina Patel*[1], *Zhengjun Yue*[1], *Odette Scharenborg*[1]

[1]Multimedia Computing Group, Delft University of Technology, the Netherlands

ABSTRACT

Automatic speech recognition (ASR) should serve every speaker, not only the majority "standard" speakers of a language. In order to build inclusive ASR, mitigating the bias against speaker groups who speak in a "non-standard" or "diverse" way is crucial. We aim to mitigate the bias against non-native-accented Flemish in a Flemish ASR system. Since this is a low-resource problem, we investigate the optimal type of data augmentation, i.e., speed/pitch perturbation, cross-lingual voice conversion-based methods, and SpecAugment, applied to both native Flemish and non-native-accented Flemish, for bias mitigation. The results showed that specific types of data augmentation applied to both native and non-native-accented speech improve non-native-accented ASR while applying data augmentation to the non-native-accented speech is more conducive to bias reduction. Combining both gave the largest bias reduction for human-machine interaction (HMI) as well as read-type speech.

***Index Terms*—** Speech recognition, bias mitigation, non-native accents, data augmentation, voice conversion

1. INTRODUCTION

Automatic speech recognition (ASR) systems should serve every group of speakers, not only the majority "standard" or "norm" speakers of a language, i.e., adult, typically highly-educated, first-language speakers of a standardized language variety, without a speech disability. In order to build inclusive ASR systems [1], mitigating the bias against speaker groups who speak in a "non-standard" way, which we refer to as "diverse" speech and includes, e.g., children, older adults, speakers with non-native and regional accents, and speakers with a voicing disorder, is crucial.

Although no one definition of bias exists, it is often instrumentalized as the (relative) difference in word error rates (WERs) between two speaker groups, e.g., [2, 3, 4, 5]. While research into bias quantification and mitigation is only nascent, a growing body of results shows bias against speaker groups with different sociolinguistic backgrounds, including gender [6, 7, 8], non-native accents [9, 10, 11], age [3, 12, 13, 14], and regional speech variety [2, 15, 16, 17, 18].

Researchers have been looking into different techniques to mitigate these biases. For instance, [9] compared different types of data augmentation in a state-of-the-art (SotA) hybrid model, specifically speed and volume perturbation [19] and pitch shift [20], and used transfer learning, specifically fine-tuning [21] and multi-task learning [22], as methods to reduce biases against child and adult non-native speakers of Dutch. They found that speed perturbation and pitch shift are beneficial to bias reduction, while the transfer learning techniques and volume perturbation were not. Similarly, [10] examined the effects of using speed-perturbed data in addition to synthetic non-native-accented speech generated using a newly developed cross-linguistic voice conversion (VC) approach [23], and compared fine-tuning and domain adversarial training (DAT) [24] in a SotA end-to-end (E2E) model. The inclusion of augmented and synthetic speech data resulted in lower WERs for both native and non-native-accented speech, as well as a non-nativeness bias reduction. Additionally, speed pertubation and voice transformation were shown to improve the recognition of non-native-accented English [25]. Data augmentation can thus help bias mitigation [9, 10, 25].

Different data augmentation techniques create different types of artificial data based on their adaptation of the original speech. In this paper, we divide these different techniques into four categories: 1) Warping features and masking part of the training speech signal to make the acoustic model more robust, e.g., SpecAugment [26]; 2) Increasing the quantity of the training data by creating artificial data that is, e.g., slower/faster or louder/softer than the original speech, e.g., using signal speed and volume perturbations [19, 27]; 3) Adding "more speakers" with the same articulation patterns as the training data, e.g., by shifting the pitch, e.g., pitch perturbations [20]; 4) Adding "more speakers" with new articulation patterns thus increasing the amount of variability in the training data, e.g., through voice conversion (VC) [23, 28, 29]. Different data augmentation techniques have different effects on recognition performance and bias reduction and these effects are dependent on the ASR architecture [9, 10]. There are many open questions regarding the role of data augmentation in improving low-resource, diverse speech recognition performance and bias reduction.

In this study, we further and systematically investigate the 1) effect of different types of data augmentation of 2) native

979-8-3503-0689-7/23/$31.00 ©2023 IEEE

Flemish and diverse, non-native-accented Flemish Dutch data separately and together, for bias reduction against non-native-accented Flemish [30] 3) for read and human-machine interaction (HMI) speech in an E2E model, with the further aim to 4) understand the relationship between different types of data augmentation and the types of diverse speech. Bias reduction is defined as reducing the WER gap between native and non-native speakers while maintaining recognition performance for native speakers.

2. METHODOLOGY

This section outlines the used datasets, experimental setup, applied data augmentations, and evaluation metrics. All models were trained on Netherlandic and/or Flemish Dutch "norm speech" from the Spoken Dutch Corpus (Section 2.1.1). We also used native and non-native-accented Flemish "diverse" speech from the JASMIN-CGN corpus for training and testing purposes (Section 2.1.2). Additionally, we used English speech data from the VCTK corpus [31] (Section 2.1.3) to generate new "non-native" accented speech. We created a baseline for Flemish. Subsequently, we ran two parallel sets of experiments, one in which different data augmentation techniques were applied to the native Flemish data and one in which these data augmentation techniques were applied to the non-native data to investigate the effect of adding more data and adding more data similar to the diverse data for which recognition performance and bias needs to be improved.

2.1. Datasets

2.1.1. Spoken Dutch Corpus (CGN)

The Spoken Dutch Corpus (*Corpus Gesproken Nederlands*, CGN) [32] contains Dutch speech data spoken by 18- to 60-year-old native speakers of Dutch from both the Netherlands (NL) and Flanders (FL). The type of speech ranges from lectures, read speech, and conversational telephone speech (CTS) to broadcast news (BN). The total amount of raw audio recordings in the training set is about 900h. After cutting the full audio recording into small chunks and removing the silent chunks, we obtain the 690.45h training set from the CGN corpus, denoted as CGN-NL-FL, which consists of 424.55h of Netherlandic Dutch data (CGN-NL) and 265.9h of Flemish (CGN-FL). To test our ASR systems on "norm" native Flemish speech, we used two Flemish test sets from CGN, i.e., a BN test set (0.4h) and a CTS test set (1.8h). The same pre-processing steps are applied to the two test sets.

2.1.2. JASMIN-CGN

The JASMIN-CGN corpus [33] contains Dutch (40.48h) and Flemish (25.07h) speech data. Compared with CGN which only has native adult speakers, JASMIN-CGN has more speaker group variations. It contains five speaker groups: (1) native children between the ages of 7 and 11 years, (2) native youngsters between 12 and 16 years, (3) non-native children between 7 and 16 years, (4) non-native adults between 18 and 60 years, and (5) native older adults over 60 years. The Flemish non-native speakers were mostly Francophones. This paper aims to quantify and mitigate the bias against non-native accents in Flemish ASR, so we only used the Flemish data of the JASMIN-CGN corpus, denoted by J-FL.

J-FL contains two speaking styles (read speech and HMI speech) spoken by native (N) and non-native (NN) speakers. We split J-FL into a training set (J-FL-Train), a validation set (J-FL-Valid), and 4 test sets (HMI-N/NN: native/non-native HMI speech, Read-N/NN: native/non-native read speech). The J-FL-Train set consists of 2 subsets: native and non-native training sets, denoted by J-FL-N and J-FL-NN. Table 1 provides the details of the J-FL data split in terms of duration (in hours), binary gender (male (M) and female (F)), and the number of speakers and utterances.

Table 1. Amount of speech data, binary gender distribution, and number of speakers and utterances for the J-FL training, validation, and test sets splits.

Name	Dur (h)	F	M	Spk	Utterances
J-FL-Train	20.20	96	81	177	35959
∟ J-FL-N	11.40	58	49	107	20573
∟ J-FL-NN	8.80	38	32	70	15386
J-FL-Valid	1.00	96	81	177	1798
HMI-N	0.58	9	9	18	1095
HMI-NN	0.71	6	6	12	1240
Read-N	1.52	9	9	18	2698
Read-NN	1.08	6	6	12	1823

2.1.3. VCTK Corpus

The VCTK corpus [31] is a non-parallel English corpus consisting of studio-quality speech from 109 native English speakers (both female and male) with various regional accents. These include accents from England, Scotland, Wales, and Northern Ireland. It contains approximately 44h of read speech. The VCTK corpus is used to train the VC models and generate non-native-accented Flemish speech data.

2.2. Data Augmentation

SpecAugment. SpecAugment [26] aims to bring more speech variability to the training data through time warping and time/frequency masking of the input features [34]. The online augmentation settings in ESPnet [34] were adopted.

Speed Perturbation. Speech perturbation [27] involves the process of sampling the original unprocessed speech signal again, leading to a modified time signal with a distorted

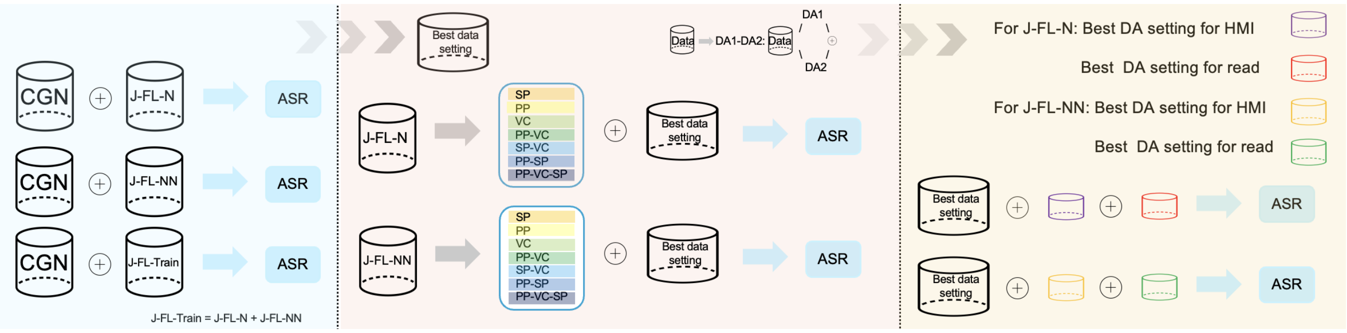

Fig. 1. The experimental pipeline. The plus sign indicates the concatenation of the data sets (left panel). The best model in terms of lower bias/WERs is used for the 7 data augmentation (DA) experiments (see Section 2.3.3) applied to the native and non-native data separately, for read and HMI speech separately (middle panel). The best DA methods for the native and non-native data are combined for the 2 speech styles separately in the final experiment (right panel; see Section 2.3.4).

tempo. We used a two-fold speed perturbation through the *sox* command with 0.9 and 1.1 perturbation factors.

Pitch Perturbation. Pitch shift changes the voice of the speaker by shifting the original speech's pitch by "cents", i.e., 1/100th of a semitone, thus increasing the speaker diversity in the training data [19]. For the speech data spoken by each speaker in the pre-augment data, two random pitch values from ranges 50-200 and 200-350 were selected to generate two new speakers for each of the original speakers using *sox*, resulting in two-fold pitch perturbation data augmentation.

Cross-lingual VC-based data augmentation. AGAIN-VC [23] was adopted for the cross-lingual VC experiments. AGAIN-VC is a SotA autoencoder-based and non-parallel VC model that disentangles the speaker and content information in the input speech data. In voice conversion, generally speaking, target speakers provide the speaker information, while source speakers provide the content information. Since we want to create non-native-accented Flemish, we used AGAIN-VC cross-lingually with accented speech data in English from the VCTK corpus as the target speech and the native/non-native-accented speech in J-FL-N/NN as the source. We used the VC technique to augment the non-native Flemish speech data in two ways resulting in new non-native-accented speech data in Flemish, i.e., using the native Flemish speakers in J-FL-N and the non-native Flemish speakers in J-FL-NN as the source speech, in order to investigate the effect of the source data on bias mitigation. The demos[1] of generated VC data are given. The pipeline for generating cross-lingual VC-augmented data is listed below:

1. **Train VC models**: Following the training setup in [23], the VCTK corpus (target) and J-FL-N/NN (sources) were used to train two separate VC models.

2. **Calculate speaker embeddings**: a pre-trained Conv-GRU[2] speaker embedding model was used to obtain the speaker embeddings in the VCTK corpus, J-FL-N, and J-FL-NN.

3. **Select source and target speakers**: The cosine speaker similarity method [35] was used to compute the speaker similarity scores between each source speaker (from J-FL-N or J-FL-NN) and each target speaker candidate (from VCTK corpus). The two target speakers from VCTK with the two highest cosine similarity values with the source speaker were used as source speakers resulting in two-fold data augmentation.

4. **Generate VC data**: Taking the selected source and target speaker from step 3, the VC models trained in step 1 were used to generate non-native-accented data from native/non-native Flemish speakers.

2.3. Experiments

First, a solid Flemish ASR baseline was constructed (Section 2.3.1). Subsequently, as depicted in Figure 1, we conducted three blocks of experiments, i.e., adding native-/non-native-accented Flemish to the training data (left panel; Section 2.3.2), carrying out data augmentation on the native- and non-native-accented Flemish separately (middle panel; Section 2.3.3), and combining the best data augmentation methods of native- & non-native-accented speech (right panel).

2.3.1. Baselines

With Flemish being a variant of Dutch, to build a SotA Flemish ASR system, we compared ASR models trained solely on Netherlandic Dutch (CGN-NL) or Flemish (CGN-FL) and on both (CGN-NL-FL). The models were evaluated on the "standard" native Flemish test sets (CTS and BN) from CGN. We selected the best-performing model on CTS and BN as our baseline.

Secondly, the HMI-N/NN and Read-N/NN Flemish test sets from J-FL are rather small. To investigate their validity

[1] https://non-native-accented-flemish.github.io/
[2] https://github.com/RF5/simple-speaker-embedding

as our test sets, we quantified the bias against the non-native-accented speech of these small test sets and compared those with the results on the full J-FL corpus (in the remainder of our experiments, the full J-FL corpus is not used, only the separate training and test sets).

2.3.2. *Adding (Target) Native/Non-native Speech Data*

To investigate the potential effect of the mismatch between the training (CGN) and test (JASMIN) data, we trained three models on CGN (depicted as a cylinder with CGN in Figure 1) to which the Flemish native (J-FL-N), non-native (J-FL-NN) or both (J-FL-Train) were added. The best-performing models for HMI-N/NN and Read-N/NN were used to conduct the subsequent data augmentation experiments. Moreover, these models are compared to those in Section 2.3.3 to untangle the effect of the specific data augmentation technique and the effect of adding more training data.

2.3.3. *Data Augmentation Experiments*

We investigated the effect of the different augmentation techniques applied to the native Flemish speech J-FL-N and the non-native-accented Flemish speech J-FL-NN separately. Moreover, the data augmentation techniques were applied to both read and HMI speech. This resulted in four sets of data augmentation experiments. For the J-FL-N experiments, the best native models for read and HMI speech from Section 2.3.2 were used as the starting point. For the J-FL-NN experiments, the best non-native-accented models from Section 2.3.2 were used as the starting point.

Speed perturbation (SP) was used to increase the variability in speech tempo in the training data which can be seen as “simply” increasing the amount of training data. Pitch perturbation (PP) was used to increase the “number of speakers” with the same accent as the original training data. The cross-lingual VC-based technique (VC) was used to add more speakers with different non-native accents compared to the original training data. The augmented data by different techniques are merged for further data augmentation (PP-VC, SP-VC, PP-SP, and PP-SP-VC).

2.3.4. *Combining Native- and Non-native-accented Data Augmentation*

To examine whether it is possible to further mitigate the non-nativeness bias in Flemish ASR, in the final block of experiments (see right panel in Figure 1), the best data augmentation settings for the native read speech and non-native read speech, respectively, were selected and combined resulting in a new model. The same was done for HMI speech. Subsequently, SpecAugment (SpecA) was then applied to these two new ASR models. For both read and HMI speech, the best data augmentation settings were chosen according to the lower bias, without hurting the ASR performance on native-accented speech.

2.4. ASR Model

The SotA ASR model is a conformer-based sequence-to-sequence model [36] in ESPnet [34]. Training and testing ran on 4 NVIDIA GeForce GTX 1080 Ti GPUs. All the ASR experiments were conducted with filterbank features.

The ASR model consists of a 12-layer conformer encoder and a transformer decoder with 5 decoder layers, all with 2048 dimensions; the attention dimension is 512 and the number of attention heads is 8. The number of batch bins is set to 10000000 for every experiment conducted in this paper. The conformer model was trained for 25 epochs using a joint connectionist temporal classification (CTC)-attention objective [37], in which the CTC and attention weights are set to 0.3 and 0.7, respectively. We set the number of iterations per epoch 10000 times. Byte Pair Encoding (BPE) units with a vocabulary size of 5000 are used as basic units. Finally, the final test model is the averaged 10-best models with 10 lowest validation losses.

2.5. Evaluation Metrics

The performance is reported in terms of WERs for the native- and non-native-accented speech and for read speech and HMI speech separately. Bias against non-native accents is conceptualized as the difference between the WER performance on the native speech and the non-native-accented speech and is calculated as follows:

$$Bias = |WER_{HMI/Read-NN} - WER_{HMI/Read-N}| \quad (1)$$

3. RESULTS

3.1. Baseline Results

Table 2 shows the baseline results, i.e., from training the ASR model solely on CGN-NL, CGN-FL, and CGN-NL-FL, respectively, for the conversational speech (CTS) and broadcast news (BN) CGN test sets. Bold indicates the lowest WER. Training on both the Netherlandic Dutch and Flemish (CGN-NL-FL) gave the lowest WER for both test sets, even outperforming the Flemish-only model from [38]. Training the ASR only on Netherlandic Dutch gave worse results than training on only Flemish, which is not surprising given the Flemish test data. We thus move forward with our experiments using the model trained on CGN-NL-FL as our baseline.

Table 3 shows the performance and bias results of the baseline model on the native and non-native speaker groups of the full JASMIN datasets and those on the JASMIN test datasets we defined for our experiments for read and HMI speech separately. The results for the Split set are slightly worse than those for the full set for read speech; however, the

HMI-NN results are considerably worse for the Split set, although the results for the native speakers are slightly better for HMI speech. Overall, we conclude that the bias mitigation task is comparable in difficulty or slightly more difficult on the Split test set compared to the the full JASMIN dataset.

Table 2. WERs of the baseline systems tested on Flemish norm speech from the CGN. Bold indicates best results.

Training data	**CTS**	**BN**
CGN-NL	22.2	52.0
CGN-FL	11.2	35.2
CGN-NL-FL	**8.9**	**32.8**

Table 3. WERs and biases of the best baseline system on the full J-FL and the split test sets from the JASMIN corpus.

	Read			**HMI**		
Test set	**N**	**NN**	**Bias**	**N**	**NN**	**Bias**
Full	**27.8**	**48.4**	**20.6**	40.0	**52.3**	**12.3**
Split	30.2	51.8	21.6	**37.8**	61.2	23.4

3.2. Data Augmentation Results

Table 4 shows the recognition results and bias of the experiments in five blocks: 1) the baseline (same as in Table 3), the baseline system augmented with SpecAugment; 2) the experiments with native and non-native-accented Flemish added; 3) native data augmentation experiments; 4) non-native data augmentation experiments; and 5) the combined system.

Applying SpecAugment. Applying SpecAugment improved recognition performance for both speaker groups and both speaking styles, but at the cost of an increased bias.

Adding natural native and non-native data. First, adding native, non-native-accented, or both to the CGN data (see block 2 of Table 4) improves recognition performance substantially (compared to Base w. SpecAug). Adding the non-native-accented data gave better recognition results than adding only native data and also substantially reduced bias against non-native-accented speech. Second, the best results for HMI-N, R-N, and R-NN were obtained when both native and non-native Flemish data from the JASMIN-CGN corpus, i.e., J-FL-Train, were added to the training data.

Native data augmentation. Comparing the results of the seven data augmentation techniques against the Base + J-FL-Train results (block 3) showed that adding different types of augmented native-accented speech data individually improved recognition performance only marginally for read and HMI speech. However, for HMI speech, adding SP, PP, or VC increased the HMI bias by 1.5%. The best results, i.e., lowest WER for bias against non-native-accented speech, for read speech were obtained when combining PP-VC (blue row) and for HMI speech when combining PP-SP (green row).

Table 4. Results of the data augmentation experiments. The shaded rows denote the lowest non-nativeness bias for read speech (blue) and HMI speech (green) in the native- and non-native-accented experiments. Best-R denotes the combination of the augmented settings of the two blue rows; Best-H denotes the combination of the augmentation settings of the two green rows.

	Read			**HMI**		
Training Data	**N**	**NN**	**Bias**	**N**	**NN**	**Bias**
Base	30.2	51.8	21.6	37.8	61.2	23.4
Base w. SpecA	22.0	46.2	24.2	33.8	59.1	26.3
N/NN Data Addition						
Base + J-FL-N	7.2	25.4	18.2	**19.8**	39.7	19.9
Base + J-FL-NN	10.0	12.5	**2.5**	24.1	21.7	2.4
Base + J-FL-Train	**6.1**	**11**	4.9	20.5	**20.7**	**0.2**
Data Augmentation on J-FL-N						
(N) SP	4.9	10.0	5.1	19.5	20.2	1.7
(N) PP	5.0	9.8	4.8	18.3	19.9	1.6
(N) VC	4.8	10.1	5.3	19.0	20.7	1.7
(N) PP-VC	4.2	**8.9**	**4.7**	18.6	19.7	0.9
(N) SP-VC	4.2	9.1	4.9	**17.9**	20.6	2.7
(N) PP-SP	4.6	9.5	4.9	18.5	**19.3**	**0.8**
(N) PP-VC-SP	**4.0**	9.0	5.0	18.4	20.5	1.9
Data Augmentation on J-FL-NN						
(NN) SP	5.7	9.4	3.7	19.2	19.3	**0.1**
(NN) PP	5.7	9.3	3.6	19.0	**18.9**	**0.1**
(NN) VC	5.3	9.4	4.1	19.4	20.1	0.7
(NN) PP-VC	5.1	8.5	3.4	19.8	20.0	0.2
(NN) SP-VC	4.9	8.4	3.5	**18.4**	19.2	0.8
(NN) PP-SP	5.3	8.6	3.3	18.7	20.0	1.3
(NN) PP-VC-SP	**4.8**	**8.0**	**3.2**	19.2	19.5	0.3
Combining N and NN Data Augmentation						
Best-R	3.9	7.5	3.6	19.0	18.4	**0.6**
Best-R w. SpecA	**3.7**	**6.6**	**2.9**	**18.7**	**17.0**	1.7
Best-H	4.5	8.6	4.1	17.6	18.9	1.3
Best-H w. SpecA	**3.4**	**6.9**	**3.5**	**17.5**	**17.4**	**0.1**

Non-native data augmentation. Comparing the results of the seven data augmentation techniques against the Base + J-FL-Train results showed that adding different types of augmented non-native-accented speech data individually improved recognition performance again marginally for both read and HMI speech. PP augmentation yielded the lowest WER for the non-native-accented speech for HMI speech and the lowest bias. Combining all data augmentation techniques gave the best recognition and bias results for read speech.

Combining the best native and non-native data augmentation approaches. Combining the best augmentation approach for the native- and non-native-accented speech for read speech (blue rows; Best-R) further reduced the recognition performance and bias for read speech and for the non-native-accented HMI speech and bias. Adding SpecAugment gave the overall best results with the lowest recognition rates for both N and NN speech, and with the smallest bias for read speech at the cost of a slight increase in the bias for HMI. Combining the best augmentation approaches for HMI speech (green rows; Best-H) also improved recognition performance for both speaker groups and speech types. Adding SpecAugment further improved both recognition performance and bias. Interestingly, the models that combined the best augmentation strategies for read speech and HMI speech, respectively, also improved recognition performance for the speech type on which it was not optimized. This suggests that adding more artificial data helps in both recognition performance and bias reduction.

A comparison across all models shows that the best results were obtained when adding data through multiple augmentation techniques. We thus show that pitch perturbations (PP) and speed perturbations (SP) of native- and non-native-accented data, and SpecAugment, where also adding native- and non-native-accented voice-converted data improved native and non-native-accented speech recognition and reduced bias against non-native-accented speech.

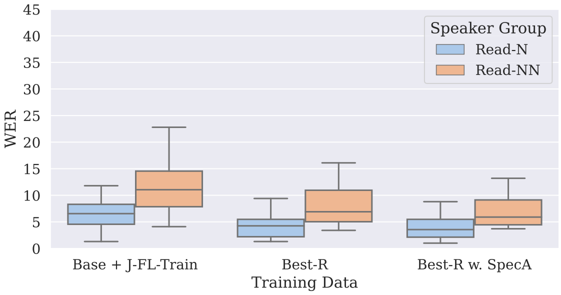

Fig. 2. Distributions of WERs for Read-N/NN.

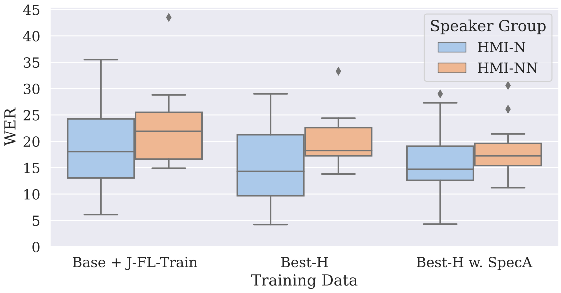

Fig. 3. Distributions of WERs for HMI-N/NN.

Figures 2 and 3 show the distributions of the WERs for the nine native and six non-native speakers from our test sets across selected models, i.e., the model trained on (CGN-NL-NL and J-FL-Train) and the respective best models w/o. SpecA. They illustrate how the models with augmented data are not only able to reduce the overall WERs and biases, but also decrease the size of the WER distributions.The results provided in Table 4 show that there was no one data augmentation method/combination of data augmentation methods that gave the lowest WERs/bias across the board. However, the lowest bias for read and HMI speech can be achieved separately in two ASR models. Furthermore, we showed that more data always resulted in better performance.

4. DISCUSSION AND CONCLUSION

In this study we systematically investigated the effect of different types of data augmentation and the type of data it was applied to for non-nativeness bias reduction. The experimental results indicate that, overall, adding more data, whether natural (see Block 2 of Table 4) or artificial (Blocks 3-5), helps reduce WERs for native- and non-native-accented speech and the non-nativeness bias. Interestingly, adding artificial non-native data created using native speakers also improves non-native-accented speech recognition (Block 3, VC), but at the cost of a small increase in bias when compared to a model trained on a small amount of natural non-native-accented speech (Block 2). If a small amount of non-native-accented speech data is available, it is preferable to use that for data augmentation as adding artificial non-native-accented data slightly outperforms adding native artificial data. The type of data augmentation does not really seem to matter, as all techniques reduced the bias and improved the performance, though different combinations do yield slightly different results. This paper extends the growing body of work investigating non-native speech recognition, e.g., [25, 11]. Specifically, we expand on the findings for bias mitigation against non-native-accented Dutch [9, 10]. We not only showed that SP, PP, and VC helped reduce bias against non-native-accented Flemish, found in [30], but also showed that PP outperforms SP [9]. Comparably to [25], we found that SP outperforms VC [10]. This suggests that adding "more speakers" might be more beneficial than adding more data. The lower results for VC might be due to the quality of the generated speech. Future work may look into the effect of VC quality on bias reduction.

In conclusion, we found that the addition of augmented data can considerably reduce WERs and biases. With SpecAugment, adding more speed-perturbed native speech data and more speakers with the same accents as both native and non-native speakers to the training data nearly removed the bias against non-native accents for HMI speech in a SotA Flemish ASR system; adding more speakers with native and (new) non-native accents and increasing the amount of non-native-accented speech reduced the bias against non-native-accented speech for read speech from 21.6% to 2.9%. Applying non-native-accented speech data augmentation always led to bias reduction while combining it with native speech data augmentation further improved recognition performance without enlarging the bias. We advocate for more research on bias and specifically bias mitigation in ASR.

5. REFERENCES

[1] O. Scharenborg, "Inclusive speech technology," https://www.tudelft.nl/tu-delft-safety-security-institute/events/webinars/webinar-inclusive-speech-technology-1.

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

[3] T. Patel and O. Scharenborg, "Using data augmentations and vtln to reduce bias in dutch end-to-end speech recognition systems," *arXiv preprint arXiv:2307.02009*, 2023.

[4] Z. Liu, I. Veliche, and F. Peng, "Model-based approach for measuring the fairness in asr," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6532–6536.

[5] P. DHERAM, M. Ramakrishnan, A. Raju, F. I-Chen, B. King, K. Powell, M. Saboowala, K. Shetty, and A. Stolcke, "Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities," in *Proc. Interspeech*, 2022, pp. 1268–1272.

[6] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the first ACL workshop on ethics in natural language processing*, 2017, pp. 53–59.

[7] M. Garnerin, S. Rossato, and L. Besacier, "Gender representation in French broadcast corpora and its impact on ASR performance," in *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, New York, NY, USA, 2019, AI4TV '19, pp. 3–9, Association for Computing Machinery.

[8] M. Garnerin, S. Rossato, and L. Besacier, "Investigating the impact of gender representation in ASR training data: a case study on Librispeech," in *3rd Workshop on Gender Bias in Natural Language Processing*, Online, France, Aug. 2021, pp. 86–92, Association for Computational Linguistics.

[9] Y. Zhang, Y. Zhang, Patel T., and O. Scharenborg, "Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems," in *Proc. 1st Workshop on Speech for Social Good (S4SG)*, 2022, pp. 15–19.

[10] Y. Zhang, Y. Zhang, B. M. Halpern, T. Patel, and O. Scharenborg, "Mitigating bias against non-native accents," in *Proc. Interspeech*, 2022, pp. 3168–3172.

[11] S. Wills, Y. Bai, C. Tejedor-Garcia, C. Cucchiarini, and H. Strik, "Automatic speech recognition of non-native child speech for language learning applications," in *12th Symposium on Languages, Applications and Technologies (SLATE)*, 2023.

[12] T. Pellegrini, I. Trancoso, A. Hämäläinen, A. Calado, M. S. Dias, and D. Braga, "Impact of age in ASR for the elderly: preliminary experiments in European Portuguese," in *Advances in Speech and Language Technologies for Iberian Languages*, pp. 139–147. Springer, 2012.

[13] A. Hämäläinen, A. Teixeira, N. Almeida, H. Meinedo, T. Fegyó, and M. S. Dias, "Multilingual speech recognition for the elderly: The AALFred personal life assistant," *Procedia Computer Science*, vol. 67, pp. 283–292, 2015.

[14] H. K. Kathania, S. Reddy Kadiri, P. Alku, and M. Kurimo, "Study of formant modification for children ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7429–7433.

[15] M. Sawalha and M. Abu Shariah, "The effects of speakers' gender, age, and region on overall performance of arabic automatic speech recognition systems using the phonetically rich and balanced modern standard arabic speech corpus," in *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2*, 2013.

[16] M. Ngueajio and G. Washington, "Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review," in *International Conference on Human-Computer Interaction*, 2022, pp. 421–440.

[17] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. 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.

[18] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, and M. Wieling, "Making more of little data: Improving low-resource automatic speech recognition using data augmentation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, July 2023, pp. 715–729, Association for Computational Linguistics.

[19] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, and Y. Wang, "Data augmentation for children's speech recognition–the" ethiopian" system for the slt 2021 children speech recognition challenge," *arXiv preprint arXiv:2011.04547*, 2020.

[20] C. Bellettini and G. Mazzini, "Reliable automatic recognition for pitch-shifted audio," in *Proceedings of 17th International Conference on Computer Communications and Networks*. IEEE, 2008, pp. 1–6.

[21] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," *arXiv preprint, arXiv:1609.03193*, 2016.

[22] Z. Chen, S. Watanabe, H. Erdogan, and J. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[24] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[25] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, "Data augmentation improves recognition of foreign accented speech.," in *Interspeech*, 2018, number September, pp. 2409–2413.

[26] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[27] 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.

[28] B. Chen, Z. Xu, and K. Yu, "Data augmentation based non-parallel voice conversion with frame-level speaker disentangler," *Speech Communication*, vol. 136, pp. 14–22, 2022.

[29] L. Prananta, B. M. Halpern, S. Feng, and O. Scharenborg, "The effectiveness of time stretching for enhancing dysarthric speech for improved dysarthric speech recognition," in *Proc. Interspeech*, 2022, pp. 36–40.

[30] A. Herygers, V. Verkhodanova, M. Coler, O. Scharenborg, and M. Georges, "Bias in Flemish automatic speech recognition," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, 2023, pp. 158–165.

[31] Y. Junichi, V. Christophe, and M. Kirsten, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019.

[32] N. Oostdijk, "The Spoken Dutch Corpus. Overview and first evaluation," in *Proc. 2nd International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May 2000, European Language Resources Association (ELRA).

[33] C. Cucchiarini, H. Van hamme, O. van 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," in *Proc. 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006, European Language Resources Association (ELRA).

[34] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

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

[36] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[37] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017, pp. 4835–4839.

[38] B. Van Dyck, B. BabaAli, and D. Van Compernolle, "A hybrid asr system for southern dutch," *Computational Linguistics in the Netherlands Journal*, vol. 11, pp. 27–34, Dec. 2021.