Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems

Zhang, Yixuan; Zhang, Y.; Patel, T.B.; Scharenborg, O.E.

DOI
10.21437/S4SG.2022-4

Publication date
2022

Document Version
Final published version

Published in
Proc. 1st workshop on speech for social good (S4SG)

1st Workshop on Speech for Social Good (S4SG)
24-25 September 2022, Incheon, Korea

Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems

Yixuan Zhang, Yuanyuan Zhang, Tanvina Patel and Odette Scharenborg

Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands
(y.zhang-96,y.zhang-88)@student.tudelft.nl, (t.b.patel,o.e.scharenborg)@tudelft.nl

Abstract

One important problem that needs tackling for wide deployment of Automatic Speech Recognition (ASR) is the bias in ASR, i.e., ASRs tend to generate more accurate predictions for certain speaker groups while making more errors on speech from other groups. We aim to reduce bias against non-native speakers of Dutch compared to native Dutch speakers. We investigate three different data augmentation techniques - speed and volume perturbation and pitch shift - to increase the amount of non-native accented Dutch training data, and use the augmented data for two transfer learning techniques: model fine-tuning and multi-task learning, to reduce bias in a state-of-the-art hybrid HMM-DNN Kaldi-based ASR system. Experimental results on Dutch read speech and human-machine interaction (HMI) speech showed that although individual data augmentation techniques did not always yield an improved recognition performance, the combination of all three did. Importantly, bias was reduced by more than 18% absolute compared to the baseline system for read speech when applying pitch shift and multi-task training, and by more than 7% for HMI speech when applying all three data augmentation techniques during fine-tuning, while improving recognition accuracy of both native and non-native Dutch speech.

Index Terms: Automatic speech recognition, bias, transfer learning, data augmentation

1. Introduction

Automatic Speech Recognition (ASR) systems aim to deliver objective transcriptions of speech, and in order to do so they have to deal with the highly ambiguous nature and variability of human speech. However, state-of-the-art (SOTA) ASR systems do not have an equally good recognition accuracy for every speaker: Recent evidence has shown that speech variability due to, e.g., gender [1, 2, 3, 4, 5], age [6, 5], speech impairment [3], regional accents [2, 5], racial disparity [7], and non-native accents [5, 4], [8] lead to a degradation in ASR performance. There are many reasons for this bias to occur, e.g., imbalanced training data sets, a mismatch between the test data and the training data, vocal characteristics of certain speaker groups, and as recently shown by [5], specific architectures and algorithms used during ASR system development lead to different types of bias. Bias can occur at the level of the language model (LM) due to a mismatch of the speaker's use of words or grammatical structures compared with the LM training data, and at the level of the Acoustic Model (AM) due to a mismatch between the speaker's pronunciation or articulation with the AM training data (e.g., [5]). For instance, for non-native speakers, the speaker's mother tongue typically affects the pronunciation of the non-native language [9]. This deviance in pronunciation of specific sounds causes the speech recognition system's accuracy to decrease, leading to a bias against non-native listeners.

In this work, we build a Dutch SOTA ASR system and aim to reduce the bias, i.e., reduce the performance gap between non-native and native speech while not hurting performance on native speech too much. To achieve this, we focus on two approaches: increasing the amount of non-native speech data for training and using the available data more effectively while training. To increase the amount of training data, several data augmentation techniques applied in the time and/or frequency domain have been tried and proven to be successful for ASR, including for the recognition accuracy of foreign accented speech [10]. In this paper, we use speed perturbation [11], volume perturbation [12] and pitch shift [13] to augment the speech data. Next, to use the data more effectively when only limited training data of the target speech is available (as in our case only limited non-native accented Dutch speech is available), transfer learning can be used to use the available data in a more effective way. Transfer learning aims to transfer the knowledge obtained from the source or in-domain data, in our case native Dutch, to learn the unknown characteristics of the target or out-of-domain data, i.e., non-native accented Dutch. We investigate two types of transfer learning: fine-tuning and multi-task learning, both of which have proven effective in reducing the word error rate (WER) of the target speech when compared with models trained with source data only [14, 15]. Fine-tuning, though, might result in a deterioration of the recognition performance of the source speech as it is not concerned with the source speech. Multitask learning, on the other hand, aims to achieve good performance on both the source and target speech.

We analyse the ASR recognition performance of native and non-native speech and the bias for both read and spontaneous human-machine interaction (HMI) speech. Based on the previous research in [5], which showed that a hybrid Time-Delay Neural Network Factorisation (TDNNF)-Hidden Markov Model (HMM) showed a smaller bias against non-native speaker groups in comparison to End-to-End (E2E) models, we aim to reduce bias against non-native accented Dutch using the TDNNF-HMM-based ASR model and compare the three data augmentation techniques and two transfer learning techniques outlined above. This research complements our recent work in which we investigated reducing bias against non-native speech in E2E models using voice conversion (VC) and speed perturbation based data augmentation techniques and using fine-tuning and domain-adversarial training (DAT). The results of the current paper contribute to the study on the fairness of inclusive ASR systems by demonstrating that data augmentation and more effective use of the available training data can reduce bias against non-native Dutch speakers.

2. Methodology

This section discusses the database details (Section 2.1) and the baseline hybrid TDNNF-HMM model (Section 2.2). The

10.21437/S4SG.2022-4

three data augmentation techniques (Section 2.3) and the training techniques (Section 2.4) and the experiments and evaluation method (Section 2.5) are also described here.

2.1. Corpora

2.1.1. The spoken Dutch corpus (CGN)

The CGN is a dataset of contemporary standard Dutch as spoken by adults (age 18-approximately 60 years) in the Netherlands and Flanders [16]. It covers different speaking styles including read, broadcast news (BN), and conversational telephone speech (CTS). The size of the corpus is close to ten million words (about 1,000 hours of speech), two thirds of which originates from the Netherlands and one third from Flanders [16]. Only the data from the Netherlands was used in our study. We followed the standard training and test partitions as defined by [17]. The total amount of training data is 423 hours of standard Dutch speech data after segmenting the audio files to clips of at least 6 seconds in duration, and is denoted by C_{train}.

2.1.2. Jasmin-CGN corpus

The Jasmin-CGN corpus is an extension of the CGN corpus [18]. It consists of read speech and Human-Machine Interaction (HMI) speech spoken by native Dutch speakers (children: 7-11 years, teenagers: 12-16 years, older adults: above 65 years) and non-native Dutch speakers (children: 7-16 years, adults: 18-60 years) with a wide range of native languages (children: 7-16 years, adults: 18-60 years).

The training set from Jasmin-CGN contains 36.12 hours of speech data, and is denoted by J_{train}. The J_{train} is divided into 26.73 hours read speech and 9.39 hours HMI speech. The training data has 14.1 hours of non-native speech (10.42 hours read data and 3.69 hours HMI data) and 22.02 hours of native speech (16.31 hours read data and 5.70 hours HMI data). We defined four test sets with the number of female and male speakers identical in all test-sets. These test sets are the same as used in our recent work [19]:

- R_D: Native Dutch read speech data: 1.45 hours.
- R_{NN}: Non-native Dutch read speech data: 1.63 hours.
- H_D: Native Dutch HMI speech data: 0.68 hours.
- H_{NN}: Non-native Dutch HMI data: 0.36 hours.

2.2. Baseline state-of-the-art ASR system for Dutch

The baseline ASR model is a hybrid DNN-HMM system with TDNNF architecture as in Figure 1. The training is done using the Kaldi tool-kit [20] and the parameters are similar as in [5]. The AM consists of 12 TDNNF layers of dimension 1024, and was trained with the lattice-free maximum mutual information (LF-MMI) criterion for 4 epochs. 100-dimensional i-vectors were appended to the high resolution MFCC input features for speaker adaptation purpose. Context-dependent phone alignment labels used for training the AM were obtained by using a GMM-HMM trained beforehand with the same training data as that for the TDNNF. The LM trained is a tri-gram model. The baseline model was trained with C_{train} data as mentioned in Section 2.1.1.

2.3. Data augmentation techniques

The three different data augmentation techniques investigated in the paper are discussed next.

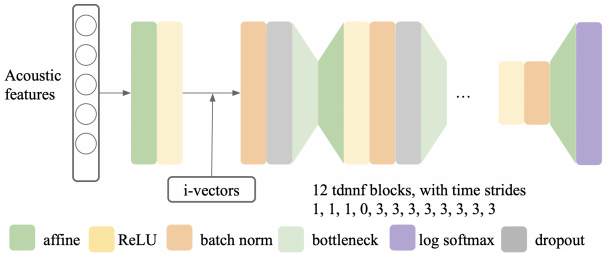

Figure 1: *The TDNNF-architecture.*

2.3.1. Speed Perturbation

Speed Perturbation re-scales the speed of the speech recordings in the time domain with a perturbation factor that changes both the audio duration and the spectral envelope [21]. The perturbation factors we applied were $\{0.9, 1, 1.1\}$. The standard Kaldi speed perturbation script [22] is used.

2.3.2. Volume Perturbation

Similar to speed perturbation, volume perturbation [21] re-scales the volume of the audio segments. The same rescaling factors as used for speed perturbation are applied for volume perturbation, $\{0.9, 1, 1.1\}$. We used the standard Kaldi script that modifies the wav.scp to perturb the volume [23].

2.3.3. Pitch Shift

The pitch shift technique allows the original pitch of a sound to be raised or lowered [24]. In our work, which uses the $librosa$ function $librosa.effects.pitch\_shift$ where the pitches of audio snippets are shifted by $\{\pm 2\}$ semitones [25]. A semitone corresponds to multiplying the number of Hertz (Hz) by $2^{\frac{1}{2}}$.

2.4. Transfer Learning

Two transfer learning techniques were investigated and compared. Later, these are also compared with the standard training method (referred to as in-domain training).

2.4.1. Fine-tuning

Fine-tuning takes the initial baseline model trained on C_{train}, and then trains the new model with a target data set (see Section 2.5). Layer transfer was employed during training, where the parameters of the layers are transferred from the baseline to be the initial values of the new model. The model is trained for four epochs, which is the same as the number of epochs used in the baseline. During fine-tuning, the baseline Gaussian Mixture Model (GMM), i-vector extractor, tree, and TDNNF architecture are used, while the target training data and a fused tri-gram language model in which the words and their combinations from both Jamin-CGN and CGN are used, as Jasmin-CGN contains phones and words unseen in CGN.

2.4.2. Multi-task Learning

During multi-task learning, the system model is trained for multiple tasks using shared information, which allows the model to exploit similarities and differences between the two tasks to create a model that is better able to generalise than models trained on a single task. Multitask learning starts from scratch by learning from CGN and Jasmin-CGN in parallel.

Figure 2 shows how multi-task learning is implemented in the AM in our TDNNF-architecture. During multi-task learning, the model is trained for recognition of the speech in CGN and recognition of the speech in Jasmin-CGN. The acoustic features and the acoustic model are shared except for the last hidden layer of the neural network in the AM. The features include 100-dimensional i-vectors extracted from the global i-vector extractor trained on both CGN and Jasmin-CGN and appended to the MFCC features. The AM is the TDNNF model and the LM is the fused tri-gram model in which the text and words from both Jamin-CGN and CGN are used.

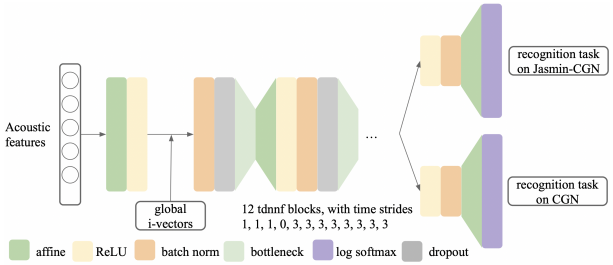

Figure 2: *Multi-task learning as implemented in the TDNNF-architecture for acoustic modeling.*

2.5. Experimental Setup and Evaluation

In our experiments only the native Dutch speech CGN C_{train} data was used to train the initial baseline model. Thereafter, to investigate the influence of adding the Jasmin-CGN data and the different types of augmented data on the training procedures, the Jasmin-CGN data and the augmented data were added to the training procedures in five ways:

- a) the original J_{train} speech data;
- b) a) + speed perturbed speech from J_{train};
- c) a) + volume perturbed speech from J_{train};
- d) a) + pitch shifted speech from J_{train};
- e) a) + b) + c) + d)

Both the native and non-native accented speech from Jasmin-CGN is used for speech, volume and pitch shift augmentation. Thereafter the training is carried out in three ways as below for each of the above five data combinations:

- In-domain training: The model is trained on the CGN C_{train} and the Jasmin-CGN and the augmented data simultaneously [i.e., data a) to e)].
- Fine-tuning: The baseline model trained on C_{train} is fine-tuned with the Jasmin-CGN data and the augmented data [i.e., data a) to e)].
- Multi-Task Learning: The model is trained on the CGN C_{train} and the Jasmin-CGN and the augmented data simultaneously [i.e., data a) to e)]. The last hidden layer and output layer are independent per data set.

All 15 models are evaluated on the four defined test sets of Jasmin-CGN in terms of the word error rate (WER). We also use bias to measure the fairness of the system. Here, Bias is defined as the absolute difference in WER between the non-native Dutch speakers and the native Dutch speakers, e.g. $B = |WER_{NN} - WER_D|$, and is calculated for read speech (B_R) and HMI (B_H) speech separately.

3. Experimental Results

This Section shows the results for the baseline and three training methods with the five different data augmented training sets.

3.1. Baseline results

Table 1 shows the recognition results in WER for the native and non-native accented speech and the bias against the non-native accented speech, for read and HMI speech separately. Row 1 of Table 1 shows the recognition and bias results of the baseline model, which was only trained on the CGN native speech, for read speech and conversational speech. Not surprisingly, read speech is better recognised than conversational speech. At the same time, the bias against non-native listeners is about twice as large for read speech compared to that for conversational speech. While the recognition performance for read and HMI speech for the non-native accented speech was relatively similar, this was not the case for the native speech: the larger bias in read speech is primarily due to a much better recognition result for native read speech compared to native HMI speech.

3.2. Data augmentation and transfer learning results

As shown in Table 1 for in-domain training the results on adding Jasmin-CGN and the perturbed data with different augmentation techniques, decreases the WER and bias significantly as compared to the baseline. Within the in-domain experiments, we observe that the different data augmentation techniques when applied alone give only little improvement in recognition performance for both the native and non-native speakers (and a small deterioration when only volume perturbed data is applied). Applying all three data augmentation techniques, however, leads to a reduction in WER and the lowest WER for both the native and the non-native speaker groups for both read speech and HMI speech. The lowest bias for HMI is obtained when using all three data augmentation techniques, while for read speech the lowest bias was observed when volume perturbed data was added, but this is due to an increase in WER for the native speech which was larger than the increase in WER for the non-native accented speech. In general, read speech is better recognised that HMI speech, which is true for all speaker groups irrespective of the data augmentation techniques applied.

When fine-tuning is applied we observe a similar trend as for in-domain training. That is, different results are observed when different data augmentation techniques are applied alone (with an increase in WER when only volume perturbed data is applied) with the best recognition results obtained when all three data augmentation techniques are applied for both native and non-native speakers and both read and HMI speech. Also, with fine-tuning, the smallest bias for read speech was observed when only the Jasmin data was used, but again at the cost of high WERs for both speaker groups and both types of speech. For HMI speech, the smallest bias was observed when all augmented data was added during fine-tuning.

For multi-task training, the smallest bias for read speech is also the smallest bias overall, which is obtained with pitch-shifted Jasmin-CGN data. The smallest bias in HMI speech comes from training with speed-perturbed Jasmin-CGN. Among the techniques employed, fine-tuning and multi-task learning reduce the bias more than simply including the target non-native speech as in-domain data. By observing the best performance of each method, we can conclude that data augmentation contributed to the reduction of both WER and bias, and among all the data augmentation techniques adopted, pitch

Table 1: *WERs(%) on the read and HMI native/non-native speech of the Jasmin-CGN data for the models trained with different training methods (in-domain, fine-tuning, and multi-task) and different types of augmented speech data. SP refers to speed perturbation; VP refers to volume perturbation; PS refers to pitch shift. Column-wise, the lowest WER and bias are denoted in bold.*

Method	Datasets	Word Error Rate (WER %)				$Bias$	
		R_D	R_{NN}	H_D	H_{NN}	B_R	B_H
Baseline	C_{train}	20.80	48.04	30.90	44.57	27.24	13.67
in-domain	C_{train}, J_{train}	17.97	31.65	28.8	37.95	13.68	9.15
	$C_{train}, J_{train} + SP$	17.55	30.13	29.47	36.65	12.58	7.18
	$C_{train}, J_{train} + VP$	20.49	32.54	29.9	37.65	12.05	7.75
	$C_{train}, J_{train} + PS$	17.26	30.04	28.59	36.33	12.78	7.74
	$C_{train}, J_{train} + SP + VP + PS$	16.82	30.04	**27.95**	**34.66**	13.22	6.71
fine-tune	J_{train}	15.61	31.09	45.24	53.73	15.48	8.48
	$J_{train} + SP$	15.31	30.89	45.1	52.81	15.58	7.71
	$J_{train} + VP$	15.66	31.45	46.46	53.96	15.79	7.5
	$J_{train} + PS$	13.85	30.3	47.06	54.55	16.45	7.49
	$J_{train} + SP + VP + PS$	**12.64**	29.91	43.79	50.1	17.27	**6.31**
multi-task	C_{train}, J_{train}	21.11	34.8	29.05	35.98	13.69	6.93
	$C_{train}, J_{train} + SP$	20.03	34.05	28.67	35.37	14.02	6.7
	$C_{train}, J_{train} + VP$	20.84	33.73	29.01	35.86	12.89	6.85
	$C_{train}, J_{train} + PS$	18.79	27.88	28.29	35.06	**9.09**	6.77
	$C_{train}, J_{train} + SP + VP + PS$	17.05	**27.87**	28.03	34.99	10.82	6.96

shift is proven the most effective in most cases.

Furthermore, importantly, the lowest bias does not necessarily correspond to the lowest WER: for read speech, the model with the lowest bias has fairly good WERs (but not the lowest WERs), while for HMI speech, the WERs of the model with the lowest bias are relatively high. Looking at the WERs across different methods, we can see that multi-task learning shows the lowest bias for most datasets compared to in-domain training and fine-tuning. Multi-task learning maintains the balance between the WER and the bias the best among the methods adopted, but at the cost of slightly higher WERs on native speakers compared with fine-tuning. Overall, multi-task learning with pitch-shifted data seem to be the best choice if we aim to reduce the bias without causing performance degradation on native speakers. In all experiments we observed better WERs for read speech than for HMI speech. On the one hand, this is because of the typically clearer articulation of read speech compared to the more spontaneous HMI speech. On the other hand, we had less HMI than read speech data available.

4. General Discussion and Conclusions

In this paper, we have shown that bias against non-native speakers can be reduced substantially using a combination of different techniques. The results show that the application of data augmentation techniques reduces bias against non-native-accented speech of HMI speech more than it reduces the bias for read speech. This suggests that the recognition accuracy of the ASR system is more sensitive to the change HMI speech data. Speech of native adults is recognised better than that of non-native adults, but augmented non-native speech data makes the model fit the non-native side more, hence a performance degrade on native speech recognition in some cases.

Among the data augmentation techniques adopted, pitch shift contributed the most to the overall reduction in bias. A possible explanation could be that, compared to speaking volume and speaking speed, the pitch difference between native and non-native speakers is bigger and gives more variation to the speech data within a dataset. Another noticeable finding observed from the table is that combining all data augmentation techniques does not necessarily lead to better performance in terms of bias reduction, as sometimes training with only one set of augmented data has lower bias. As for the effect of transfer learning, the results show that the application of fine-tuning makes the model work better for read speech type of data than the HMI data. One possible reason could be that the read speech is more in quantity than HMI speech. Hence, on fine-tuning, the model is biased to read speech than the HMI speech. Hence, fine-tuning also requires that the data requirements are matched. On the other hand, multi-task learning has managed to avoid this kind of performance degradation. Furthermore, multitask learning enforces more fairness across native/non-native speaker groups than fine-tuning. Possible future work like making the hidden layers more task-specific could be beneficial when speech characteristics are very different.

In our recent work with E2E models we used voice conversion and speed perturbation based augmentation methods to reduce bias against non-native speech [19]. Although our previous work on quantifying bias against non-native accented Dutch showed better recognition performance of the TDNNF-HMM models compared to the E2E models, comparing the results on reducing bias in E2E [19] with those reported here, we observe that adding Jasmin data to the training data, reduces the WERs for E2E more than the hybrid models for in-domain training and fine-tuning with the respective augmentation techniques. Specifically, for the fine-tuning experiment with J_{train}, the E2E model achieved a WER of 5.00% on R_D, 20.42% on R_{NN}, 21.27% on H_D, and 35.26% on H_{NN}. Thus, the E2E models show stronger modeling potential, although at the expense of a usually higher bias than that of the hybrid model. To further study bias reduction for E2E and hybrid models, we plan to perform a detailed evaluation of the two using similar augmentation techniques.

5. References

[1] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *INTERSPEECH, Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2005, pp. 2205–2208.

[2] L. W. Kat and P. Fung, "Fast accent identification and accented speech recognition," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, vol. 1, 1999, pp. 221–224 vol.1.

[3] L. Moro-Velázquez, J. Cho, S. Watanabe, M. Hasegawa-Johnson, O. Scharenborg, H. Kim, and N. Dehak, "Study of the performance of automatic speech recognition systems in speakers with parkinson's disease," in *INTERSPEECH*, 09 2019, pp. 3875–3879.

[4] Y. Wu, D. Rough, A. Bleakley, J. Edwards, O. Cooney, P. Doyle, L. Clark, and B. Cowan, "See what i'm saying? comparing intelligent personal assistant use for native and non-native language speakers," in *Mobile HCI 2020*. United States: Association for Computing Machinery (ACM), Oct. 2020.

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

[6] M. Abushariah and M. Sawalha, "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 *The 2nd Workshop of Arabic Corpus Linguistics WACL-2*, 01 2013.

[7] R. Tatman and C. Kasten, "Effects of talker dialect, gender and race on accuracy of bing speech and youtube automatic captions," in *INTERSPEECH*, 08 2017, pp. 934–938.

[8] T. Tien Ping, "Automatic speech recognition for non-native speakers," Theses, Universit 'e Joseph-Fourier - Grenoble I, Jul. 2008.

[9] M. L. G. Lecumberri, M. Cooke, and A. Cutler, "Non-native speech perception in adverse conditions: a review," *Speech Communication*, vol. 52, no. 11-12, p. 864, Nov. 2010.

[10] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, "Data Augmentation Improves Recognition of Foreign Accented Speech," in *Proc. Interspeech 2018*, 2018, pp. 2409–2413.

[11] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition." in *INTERSPEECH*. ISCA, 2015, pp. 3586–3589.

[12] 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.

[13] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," 2013.

[14] Z. Chen, S. Watanabe, H. Erdogan, and J. R. 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.

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

[16] N. Oostdijk, "The spoken Dutch corpus. overview and first evaluation," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens, Greece: European Language Resources Association (ELRA), May 2000.

[17] L. van der Werff, "laurensw75/kaldi_egs_cgn." [Online]. Available: https://github.com/laurensw75/kaldi_egs_CGN

[18] 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 *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA), May 2006.

[19] Y. Zhang, Y. Zhang, B. Halpern, T. Patel, and O. Scharenborg, "Mitigating bias against non-native accents," in *INTERSPEECH*, 2022.

[20] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The kaldi speech recognition toolkit," in *IEEE workshop*, 2011.

[21] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015, pp. 3586–3589.

[22] Kaldi Recipe Used. [Online]. Available: https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/utils/data/perturb_data_dir_speed_3way.sh

[23] Kaldi Recipe Used. [Online]. Available: https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/utils/data/perturb_data_dir_volume.sh

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

[25] Librosa: Pitch Shift Function Used. [Online]. Available: https://librosa.org/doc/latest/generated/librosa.effects.pitch_shift.html