Improving Northern Regional Dutch Speech Recognition by Adapting Perturbation-based Data Augmentation

Nikolay Zhlebinkov
Supervisor(s): Tanvina Patel, Odette Scharenborg
EEMCS, Delft University of Technology, The Netherlands
22-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Automatic speech recognition (ASR) does not perform equally well on every speaker. There is bias against many attributes, including accent. To train Dutch ASR, there exists CGN(Corpus Gesproken Nederlands) and as an extension, the JASMIN corpus with annotated accented data. This paper focuses on improving ASR performance for NRAD (Northern regional accented Dutch) speech, training on speakers from the region of Overijssel. To achieve this improvement, the corpus data is augmented using Vocal Tract Length Perturbation (VTLP), which entails randomly warping the frequency of each recording using a factor in the range [0.9, 1.1]. The baseline and augmented ASR systems are trained using trigram GMM-HMM (Gaussian mixture model hidden Markov models) through the Kaldi toolkit on the DelftBlue supercomputer. This leads to improvements on word error rates (WER) for all speaker groups and styles, with an overall relative improvement of 14,64% and the biggest improvement observed for male speakers - from 25.15% WER to 19,68% WER. The impact of this augmentation on other accents and non-accented speech is not explored. This experiment can serve as a stepping stone for developing overall more robust and less biased Dutch ASR.

Index Terms: speech recognition, data augmentation, vocal tract length perturbation

1. Introduction

Automatic Speech Recognition (ASR) systems have various applications - virtual assistants, home automation systems, automatic subtitling and more. Unfortunately, ASR does not work equally well for all users - bias exists against different speech characteristics, such as style of speaking, as well as speaker age, gender, accent or language proficiency [1][2]. This paper focuses specifically on bias against accented speech. The Dutch accented training data currently available and used in this experiment is from the JASMIN corpus [3].

The JASMIN corpus is an extension of CGN (Corpus Gesproken Nederlands) and contains speech recordings, annotated based on speaker age, gender and region of accent. This paper focuses specifically on NRAD (Northern regional accented Dutch) speech, and that is all from the Overijssel region, which has more than a million residents [4]. The training data is annotated based on speaker gender and 3 age groups - children between 7 and 11, teens between 12 and 16 and elderly people above 65. All speakers in the dataset are native. Additionally, some speakers have recorded two types of speech - read speech and human-computer interaction (HCI) [5].

ASR accuracy is very dependent on the data used to train the system and there are different approaches to gaining more data for a more robust ASR. One option is manually recording varied audio samples and annotating the text corresponding to that audio, which is a difficult and very time-consuming process. An alternative is data augmentation, which entails modifying existing data, and then adding that as additional training data for the ASR.

There exist various data augmentation techniques and of particular interest are perturbations, which entail modifying speech recordings using signal processing techniques. Several options were considered before selecting one to use for this experiment. SpecAugment combines 3 different perturbation methods and operates on the log mel spectrogram of the audio [6]. SpecSwap involves swapping frequencies and time segments [7]. Frequency perturbation involves amplifying or reducing specific frequency ranges. Pitch shift perturbation involves shifting the fundamental frequency of all utterances by a certain constant value.

Vocal Tract Length Perturbation (VTLP) is the data augmentation technique explored in this paper. It entails warping the frequency of each speaker by a random value from a range, which simulates a different vocal tract than the original speaker [8]. On average children have shorter VTL (vocal tract length) than grown females, who in turn have shorter VTL than grown males [9]. VTLP has been used in various studies and shown to lead to improvement [10, 11, 12].

The goal of this paper is to document the impact of data augmentation on Dutch ASR, specifically focusing on NRAD (Northern regional accented Dutch) speech. To quantify this impact, word error rate (WER) will be used - a common metric for ASR accuracy that is calculated as follows: $\frac{S+D+I}{N}$ where S is the number of substitutions, D - the number of deletions, I - the number of insertions and N - the total amount of words in the reference [13]. The following subquestions are outlined:

- What is the WER when recognising the initial set of accented data?
- How does the VTLP augmentation impact the overall WER?
- In what way does the impact of augmentation differ for the three age groups?
- In what way does the impact of augmentation differ for male and female speakers?

This paper is split into 5 sections. Section 2 explains the methods used to prepare the data and the tools used to train the ASR. Section 3 outlines and analyses the results, also placing it in the context of other existing work in the field and mentioning what is outside of the scope of this paper. Section 4 contains a reflection on the ethical aspects of the research and its reproducibility. In section 5 the results are summarized, conclusions are stated, and suggestions are given for possible future research.

2. Methods and Experimental Setup

The process of training an ASR begins with analyzing, extracting and splitting the available speech data. Subsequently, it is outlined how the augmentation technique is applied to increase the training data available. Finally, it is outlined how the prepared data is used to train ASR systems.

2.1. Analysing the data

The speech data used to train the ASR comes from the JASMIN corpus, which was created as an extension of the Spoken Dutch Corpus (CGN) [3]. The JASMIN corpus contains accented data in the form of speech recordings and manual transcriptions of them. The recordings are annotated based on the speaker characteristics of age, gender and region of accent, with each speaker recording read speech and human-machine dialogs, also known as human-computer interaction (HCI).

For HCI, a Wizard of Oz setup [5] is used to elicit some phenomena that are known to occur in real life use cases of spoken dialogue systems and to cause problems, but which are less present when gathering training data. The list of

phenomena includes hyperarticulation, shouting, pauses, and more [3, p.138]. In read speech, the speakers are reading a written piece of text and these phenomena are not present. The JASMIN documentation outlines the split between read and HCI speech to be 50/50, however 5.36 hours of read speech are available and only 1.52 hours of HCI. This might be due to the HCI recordings containing more silence and prompts.

The focus of this research is only the Northern region and JASMIN contains only speakers from the region of Overijssel. The script used to extract only Northern regional data can be found in the GitHub repo for this paper [14]. The speakers are either male or female and fall within 3 age groups - children ages 7-11, teens ages 12-16 and older adults ages 65+. The data totals to 6.88 hours of utterances - these are the parts of the speech files that are used by the ASR, the rest is silence or HCI prompts, and is thus not usable.

2.2. Splitting the data

A split ratio of 80/20 was chosen, resulting in a baseline split that has a test set of 1.38 hours and a training set of 5.5 hours. In order to responsibly split the data and get informative results there are three requirements that need to be kept.

Firstly, since not all speakers speak for an equal amount of time, the split happens based on speaking time and not based on the number of speakers in each set. In the GitHub repo accompanying this paper [14] you can find the script used to calculate the speaking time of each speaker based on their utterances, as well as the script that outputs an 80/20 split of speakers based on that.

Secondly, different characteristics are provided for each speaker, and those need to be evenly spread in the train and test sets. Thus the speaking time of males compared to females is preserved, as is the speaking time of small children compared to teenagers compared to elderly.

Third and last, the same speaker must not be in both the test and train set. That means that if one utterance from a speaker is in the train set, all the utterances from that speaker are in the train set.

2.3. Augmentating the data

The approach used in this paper is Vocal Tract Length Perturbation (VTLP)[8] and it is necessary to outline what VTLP is and how it affects a speech recording, before diving into how it was implemented for this research. VTLP was chosen because multiple studies have shown its success in improving ASR performance[10, 11, 12], even though the studies were not executed in such a low resource setting.

As explained by N. Jaitly and E. Hinton (2013), VTLP originates from VTLN (where N stands for Normalization) - a technique used in ASR that fits a warping factor to each speaker to remove speaker variability. The difference is that VTLP chooses warping factors randomly each time with the goal of introducing variability. The warping is applied on the Mel filter banks and warps the frequency axis such that a frequency f is mapped to a new frquency f' [8]. Warping factors below 1 correspond to frequency compression and warping factors above 1 correspond to frequency expansion [15].

Fitch and Giedd (1999) outline that overall children have shorter VTL than grown females, who in turn have shorter VTL than grown males. In the general case, a child's vocal tract length increases as they age, and during puberty this rate changes between males and females [9]. Since the average normalizing warping factor among males is higher than that among females[15], the expectation is that that lower factors result in the speaker sounding more male and grown up, while a higher warp factor results in the speaker sounding more female and child-like. However, further research would need to be done to verify this.

To execute the VTLP augmentation for JASMIN data, open-source code from nlpaug[16] was modified and used with a warp factor range=[0.9, 1.1], zone=[0, 1] and coverage=1. The same factor range was used by the originators of VTLP [8]. The sample rate of the data, both original and augmented, is 16 kHz. The warp factors that were randomly generated for the results shown in this paper are available in the GitHub repo along with the modified code for executing VTLP [14].

2.4. ASR training toolkit

To train the ASR systems in this experiment, a hybrid approach is used that separately trains and optimizes 3 components - acoustic model, language model, and lexicon model [17, 18]. The alternative is end-to-end models but that requires more training data. The acoustic model is GMM-HMM (Gaussian mixture model hidden Markov models) based, with the alternative of deep neural networks (DNN) shown to be more successful [11, 19] but also requiring more data. The language model is trigram.

In order to reliably train ASR, the Kaldi toolkit [20] is used on the supercomputer DelftBlue[21], that can be accessed remotely through a secure connection. The author of this paper has no previous background in speech recognition and scripts were provided to by the supervisor Tanvina Patel to validate the data setup and put the ASR training in motion.

3. Results and Discussion

The goal of this research is to document the impact of VTLP augmentation on NRAD (Northern regional accented Dutch) ASR. A baseline is established and compared with the augmented results in table 1, showing that the train size was doubled and the WER was reduced. In table 2 it is observed that WER was reduced for all speaker groups, with the highest baseline WER and biggest relative decrease observed for male and children speakers.

ASR accuracy for Dutch accented speech has also been documented by Feng et. al, with higher baseline WER observed for the Northern region. This can be expected, as that ASR was trained on the entire CGN and the results calculated separately for read and conversational speech [1]. VTLP augmentation was also executed on other regional Dutch accents by my colleagues, showing no improvement for Western accented speech[22] and a slight reduction of bias for Transitional accented speech[23]. These results are a lot lower than what was observed for Northern accented speech and this might be due to the random generation of warp factors.

ASR	Train / Test size (in hours)	Train / Test size (in utterances)	WER (in %)
Baseline	5.51 / 1.38	11160 / 2736	19.88
VTLP augmented	11.01 / 1.38	22320 / 2736	16.97

Table 1: *Comparison of data size and WER for baseline and for augmented ASR*

Group	Test length (hours)	Word Error Rate (%)		Relative Decrease (%)
		Baseline	VTLP	
Combined	1.38	19.88	16.97	14.64
Read	1.12	13.92	11.19	19,61
Conversational (HCI)	0.26	49.90	45.97	7,88
Male	0.55	25.15	19.68	21,75
Female	0.83	16.46	15.14	8,02
Children (age 7-11)	0.54	34.17	27.65	19,08
Teens (age 12-16)	0.36	6.05	5.87	2,98
Elderly (age 65+)	0.48	17.23	15.52	9,92

Table 2: *Comparison per speaker group between baseline and augmented ASR WER*

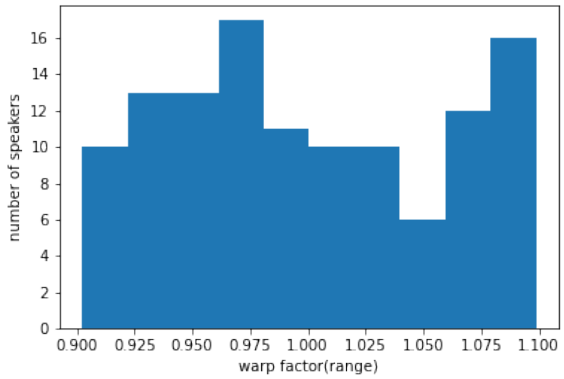

Figure 1: *Number of speakers per warp factor range*

This is one of the limitations of this paper - it is not explored how VTLP augmentation would perform on the same dataset again if different warp factors were used. Additionally, it is important to note that this ASR was trained using only NRAD data, so the impact on the augmentation for recognising other accents and non-accented speech is unknown.

The distribution of warp factors can be seen in figure 1. One theory for the improvement for male speakers might be the warp factor average being less than 1, since as is mentioned by [15], female speakers have a lower average factor for normalizing, meaning that lower factors likely contribute to a more male-like grown voice. This however would not explain the high improvement for child speakers, which on average have a shorter vocal tract than grown males [9].

Nevertheless, this paper has shown improvement in NRAD recognition for all types of speakers, with the biggest improvement for male and children speakers. Bias - the recognition gap between different speaker characteristics, can be considered reduced, because the speaker groups with highest WER also showed the biggest relative decrease in WER.

4. Responsible Research

There are three ethical concerns with this experiment. One is the reproducibility of results. Another is avoiding the addition of bias. The last is the danger of providing misleading results.

For the training of the ASR to be reproducible a step by step process has been outlined of how the data was split and how tools such as Kaldi were setup and used. For executing the augmentation, the used script and the randomly generated warp factors are available in a GitHub repo associated with this paper [14].

Bias can be observed in the way ASR systems struggle to objectively handle the large variation in speech [1]. An example of bias can be seen in table 2 in the difference in WER between male and female speakers. It can then further be seen that the bias was reduced - the gap between males and females for VTLP is smaller, as is the gap between the three age groups. The one exception is HCI speech, which is still very difficult to recognize and has improved a lot less compared to read speech. Thus one point of view can be that bias is being introduced against HCI speech by improving read speech recognition significantly more. For all other groups however, bias was reduced.

If the data is split irresponsibly it could also cause overfitting of the data and produce misleading results. To avoid this, a speaker was not allowed to be present both in the test and train set when splitting, and the train to test ratio was considered in accordance

with the low amounts of available data in order to still preserve a distribution of characteristics in both sets that is as equal as possible.

5. Conclusions and Future Work

This paper has shown that VTLP is an effective technique for improving ASR accuracy for Northern regional accented Dutch, with the biggest improvement observed for male speakers and children. Since these groups had the highest WER, this improvement has also reduced the difference in recognition between speaker characteristics, resulting in a less biased ASR. Further experiments are needed to verify what impact these augmentations have on recognizing other Dutch accents and standard non-accented Dutch. So far only GMM-HMM based ASR systems have been trained due to the low amounts of data, however possibly further augmentations of the JASMIN corpus could generate enough data to train deep neural networks (DNN), which are shown to outperform Gaussian mixed models (GMM) [11, 19].

6. Acknowledgements

Two colleagues have positively impacted this research - Alves Marinov[22] and Dragoş Bălan[23], contributing scripts for extracting the regional speakers, calculating time per speaker and generating necessary files to run kaldi, as well as providing inspiration in writing this paper. Additionally, gratitude is extended to the supervisor Tanvina Patel for guiding the team with the use of Kaldi and the available scripts, consistently answering questions and providing feedback.

7. References

[1] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," 2021. [Online]. Available: https://arxiv.org/abs/2103.15122

[2] A. Hinsvark, N. Delworth, M. Del Rio, Q. McNamara, J. Dong, R. Westerman, M. Huang, J. Palakapilly, J. Drexler, I. Pirkin, N. Bhandari, and M. Jette, "Accented speech recognition: A survey," 2021. [Online]. Available: https://arxiv.org/abs/2104.10747

[3] 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. [Online]. Available: https://aclanthology.org/L06-1141/

[4] "Key figures for the overijssel province." [Online]. Available: https://www.overijssel.nl/over-overijssel/info-overijssel/kengetallen-provincie-overijssel/

[5] N. M. Fraser and G. Gilbert, "Simulating speech systems," *Computer Speech Language*, vol. 5, no. 1, pp. 81–99, 1991. [Online]. Available: https://www.sciencedirect.com/science/article/pii/088523089190019M

[6] 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*. ISCA, sep 2019. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2019-2680

[7] X. Song, Z. Wu, Y. Huang, D. Su, and H. M. Meng, "Specswap: A simple data augmentation method for end-to-end speech recognition," in *INTERSPEECH*, 2020.

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

[9] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999. [Online]. Available: https://doi.org/10.1121/1.427148

[10] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5582–5586.

[11] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 309–314.

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

[13] "Word error rate." [Online]. Available: https://en.wikipedia.org/wiki/Word_error_rate

[14] "Scripts used in conducting the experiments for the tu delft bsc research project 2022." [Online]. Available: https://github.com/NZhlebinkov/research-project-2022

[15] L. D. Lee and R. C. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 49–60, 1998.

[16] "Python library for data augmentation in nlp." [Online]. Available: https://github.com/makcedward/nlpaug

[17] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, 2019. [Online]. Available: https://www.mdpi.com/2073-8994/11/8/1018

[18] J. Li, "Recent advances in end-to-end automatic speech recognition," 2021. [Online]. Available: https://arxiv.org/abs/2111.01690

[19] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit." IEEE Signal Processing Society, 2011, iEEE Catalog No.: CFP11SRW-USB. [Online]. Available: http://infoscience.epfl.ch/record/192584

[21] Delft High Performance Computing Centre (DHPC), "DelftBlue Supercomputer (Phase 1)," https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1, 2022.

[22] A. Marinov, "Evaluating the effect of specswap for purposes of improving wer performance of the western dutch region using the jasmin-cgn dataset," unpublished.

[23] D. Bălan, "Evaluating the use of frequency masking on a hybrid automatic speech recognizer for transitional Dutch accent of JASMIN-CGN corpus," unpublished.