Evaluating the Use of Frequency Masking on a Hybrid Automatic Speech Recognizer for Transitional Dutch Accent of JASMIN-CGN Corpus

Dragoș Alexandru Bălan
Supervisors: Tanvina Patel, Odette Scharenborg
EEMCS, Delft University of Technology, The Netherlands

June 19, 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

There are many experiments conducted with Automatic Speech Recognition (ASR) systems, but many either focus on specific speaker categories or on a language in general. Therefore, bias could occur in such ASR systems towards different genders, age groups, or dialects. But, to analyze and reduce bias, the models require significant amounts of data to be trained on, and some corpora lack that. This is where augmentation techniques can be used to generate more unique data without any further collection of it. This paper explores the use of SpecAugment's frequency masking on such a corpus, JASMIN-CGN, for the Transitional regional accent of Dutch, with a hybrid GMM-HMM architecture, in order to reduce the bias for gender or age, for this specific dialect. The experiments show that SpecAugment does not manage to lower the WER (20.8% overall compared to the baseline model, which achieves 19.5% performance), on the contrary, it even increases the bias for age. The results are mainly attributed to the combination of low amounts of data + the hybrid architecture used, which proves SpecAugment to be a useful augmentation policy only for end-to-end models.

Index Terms: ASR, hybrid ASR, speech recognition, bias, Dutch, JASMIN-CGN, audio augmentation, speech augmentation, SpecAugment

1. Introduction

Automatic Speech Recognition, or ASR, is a subfield of Computer Science and Linguistics that covers different methods and algorithms designed to take speech as audio input and recognize the words spoken, in the form of text output. Like many recognition applications, ASR systems require considerable amounts of data to be trained on in order to get high accuracy and avoid underfitting. However, collection of speech proves to be a difficult task, because first approvals from ethics committees are required, and second, a wide variety of speakers with different characteristics need to be recruited in order to avoid bias. There are plenty of corpora available for English, but other languages such as Dutch have fewer data available. That is why augmentation techniques can be used to generate more unique speech from already existing ones.

Some augmentation techniques can be more basic, such as speed perturbations, where the speech is resampled to be faster or slower, or noise perturbations, where artificial noise is added to the speech in order to make the model more robust to speech disturbances. Others, such as Vocal Tract Length Perturbation (VTLP) [1], SpecAugment [2], or SpecSwap [3], involve modifications to the spectrogram of the speech, either by warping the frequencies, masking them, or swapping them respectively and are thus more complex in their approach. Thus, the techniques can generate better or worse results due to their different approaches. Frequency masking from SpecAugment is the approach that is used by the author in this paper. The authors of SpecAugment mention that the policy improves the performance of end-to-end systems [2], but this paper will analyze its performance on a hybrid ASR system to see if similar results can be achieved. More details about this can be found in the next chapter.

It is also important to analyze and avoid bias when training an ASR system as it can often exist in most cases, as Feng et al. uncover in their work [4]. Bias mostly occurs due to imbalances of speaker distribution in the data used to train the ASR model. There have been studies on different languages which have reported that ASR systems recognize either female [5] [6] or male [7] speech better, whereas others talk about how recognizing child speech proves to be difficult due to shorter vocal tracts and higher frequencies [8] [9]. Such bias can affect the performance of ASR systems and discriminate between the different types of people, which is not desired.

Most Dutch ASR systems trained focus on the standard use of the language or oftentimes they are trained on the CGN-Dutch corpus [10] [11]. Some systems also use the JASMIN-CGN corpus [12] and focus on improving performance for specific age groups, such as the work of Pérez-Espinosa et al. for elderly speakers [13], but there is not enough research into developing a system that is unbiased for the different Dutch regional accents on the JASMIN-CGN corpus. The main regional accents observed in the corpus are West, Transitional, North, South, and Flemish. The author will focus on the Transitional region, which describes the area close to the Eastern border of the Netherlands.

For more context about the corpus that is used in this research, the JASMIN-CGN corpus [12] is an extension of the CGN-Dutch corpus [14]. It was created in order to collect speech data from Dutch children and the elderly in contrast to CGN-Dutch, which contains speech from adults only. Using this corpus, a more generalized and robust ASR system can be developed that would be less biased towards age, compared to models trained on CGN-Dutch, whilst also focusing on reducing the bias for the different regional Dutch accents. Therefore, the aim of this research paper is to answer the following main question: **Can data augmentation using SpecAugment improve the performance of an ASR system on the JASMIN-CGN corpus for the Transitional Dutch accent?** To which there are several subquestions:

- Can the Word Error Rate (WER) be lowered by augmenting data using SpecAugment for the JASMIN-CGN Transitional speech?
- Is there a significant difference in ASR performance between children, teenagers, and the elderly?
- Is there a significant difference in ASR performance between male and female Transitional Dutch speech?
- Is there a significant difference in ASR performance between read and conversational Transitional Dutch speech?

Chapter 2 of this paper will explain the methodology used to answer the question, along with a more detailed description of SpecAugment and the variation used in this paper. Chapter 3 will cover the setup of the experiments (data preparation, architecture of the model, and parameters used during the augmentation process), as well as the results obtained. Chapter 4 will summarize the paper and analyze the results obtained, as well as discuss possible improvements or questions to be answered. Chapter 5 will mention the responsible research aspects that were taken into account during research, as well as the reproducibility of the experiment. Lastly, chapter 6 mentions some acknowledgements made by the author.

2. Methodology

In order to train an ASR model, it is first needed to analyze and understand the corpus that will be used. Then, the policy of SpecAugment is covered that will be used for the augmentation of data.

2.1. JASMIN-CGN Corpus

The JASMIN-CGN Corpus [12] is an extension of the CGN-Dutch corpus [14], as mentioned in the introduction, that contains over 95 hours of Dutch speech collected from native children, teenagers, and the elderly, as well as non-native children and adults. It also aims to cover a wide variety of accents from different regions of the Netherlands (West, Transitional, Northern, and Southern) and Belgium (Flemish). There are two main speech styles collected for each speaker: read speech and conversational speech. Read speech involves the speaker reading from a script, whereas conversational, or Human-Machine Interaction (HMI), involves the speaker having a dialogue with a human-like machine.

The focus of this paper will be on the Transitional region of the Netherlands which, according to the documentation of the corpus, includes "Zeeland, Eastern Utrecht excl. the city of Utrecht, Gelders river area, incl. Arnhem and Nijmegen, Veluwe as far as the IJssel, West Friesland, and Polders". In the actual distribution of the data, only native speakers from the Gelders river area were observed to be present.

Around 10 hours of data is available for Transitional Dutch, however, the duration of speech without silence is significantly smaller. A more detailed breakdown of the data distribution can be seen in table 1.

Table 1: *Distribution of speech according to duration (in hours (h) and minutes (m))*

	Children	Teenagers	Elderly	Total
Female	1h36m	19m	1h27m	**3h22m**
Male	1h27m	26m	1h16m	**3h09m**
Total	**3h03m**	**45m**	**2h43m**	**6h31m**

As it can be seen, the data has small differences between children and the elderly, as well as males and females. Teenagers, on the other hand, have very little data available to work with, which can affect the reliability of the final results. Approximately 30% of the data is conversational, whereas the rest is read speech. Overall, due to the limited data available, it can be seen why there is a need to augment the data, which is what will be discussed next.

2.2. SpecAugment

SpecAugment [2] is an augmentation policy that combines different perturbation methods applied to the mel spectrogram of the audio. A mel spectrogram is a spectrogram that has been converted to the Mel scale, which reflects better how humans perceive sound. The perturbation methods are time warping, frequency masking, and time masking. Time warping achieves its goal by picking a random point in time from a short section of the mel spectrogram of the file and warps that point either to the left or to the right by some warping factor, in the boundaries of the chosen section. Any other parts of the spectrogram are not affected by it. The result of warping should be a spectrogram in which, for the section selected, the speech rate will be perturbed, and thus would lead to more speech generated that differs in that aspect.

Frequency and time masking, as the names suggest, involve masking a certain frequency or time range. With frequency masking, the goal is to achieve speech that sounds a bit distorted from the original, and thus make the model more robust to different speech styles or noise. Time masking, on the other hand, mutes a segment of the audio which corresponds to where the mask was applied, thus helping with training a more advanced ASR model that does not require transcripts of the words spoken.

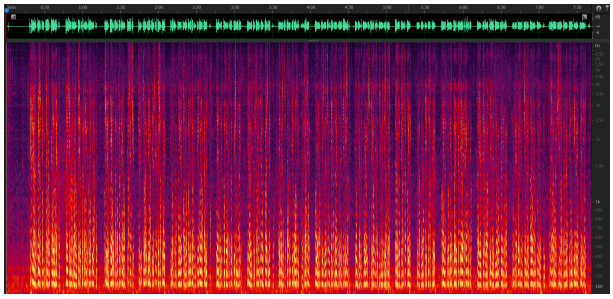

Figure 1: *Mel spectrogram of a speech file. Horizontal axis is time, vertical axis is frequency.*

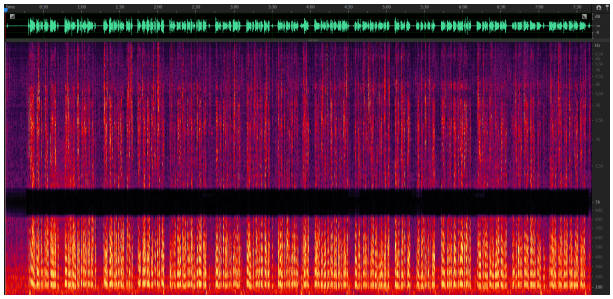

Figure 2: *Mel spectrogram of the same file, but augmented using frequency masking from SpecAugment.*

The combination of the 3 perturbation methods mentioned above is a powerful augmentation tool for generating more unique data, however, not all of them would be useful or viable to achieve in a hybrid ASR model. The authors of SpecAugment mention that time warping provides the least improvement out of the 3 methods and it is also the most computationally expensive one [2], and due to the limited time and resources available for the project, it has been decided to not implement time-warping. Time masking was also considered to be applied, however, in the case of a hybrid system, where transcripts are available, it would not make as much sense since the time-masked segments of speech would be dropped and therefore it would provide less augmented data than by applying only frequency masking for example.

Thus, from all the 3 perturbation techniques, only frequency masking was applied to the data. An example of a normal mel spectrogram can be seen in figure 1, and the application of frequency masking on the normal one can be seen in figure 2. As mentioned in the introduction, SpecAugment is a reliable augmentation policy for end-to-end models, but it is not as known how it performs for hybrid models which is what the experiment described in the next chapter will answer.

3. Experimental Setup and Results

This chapter will cover the experiments conducted and their results. First, the process of data preparation and separation into a train and a test set is discussed. Then, the technical setup is presented, including the parameters used during the augmentation process. A description of each ASR model used in the exper-

iments is provided. Lastly, the results are mentioned and the performance of the different models is analyzed.

3.1. Data preparation

The speech needs to be divided into two sets: a train and a test set. The train set is used to train the ASR model, whereas the test set is used to evaluate the model and obtain a Word Error Rate (WER) to measure the performance. In order to split the data into train and test sets in an unbiased way as well, it was first needed to not have speakers that overlap between the two sets. Then, a ratio of 80% train and 20% test data in terms of speech duration has been chosen. In the end, it has been decided to look at both gender and age and split the data for each possible combination of gender and age. Thus, the split has been made on female children, male children, female teenagers, male teenagers, female elders, and male elder people. In this way, similar distributions of characteristics between the train and test data sets were ensured. The final durations of speech used for training can be seen in table 2.

Table 2: *Distribution of speech to be used in training*

	Children	Teenagers	Elderly	Total
Female	1h17m	14m	1h10m	**2h41m**
Male	1h11m	20m	1h01m	**2h32m**
Total	**2h28m**	**34m**	**2h11m**	**5h13m**

The total amount of test data chosen amounts to 1 hour and 18 minutes, which is approx. 20% of the entire Transitional Dutch set. The speakers chosen for the test set have the following codes in the JASMIN corpus: N0000{25, 29, 38, 47, 49}, N000{160, 161}, N1000{55, 60, 63, 65}. Durations for each gender and age combination can be seen in table 3. Teenager test data is around 24% of the total available data for that age group, whereas the other categories have around 19-20% of data in the test set. This is due to the limited amount of teen speakers, which is 7 in total, so selecting speakers such that there is no overlap between the two sets and 20% duration is achieved was very difficult in this case. However, this distribution of data is considered optimal and unbiased enough to train and test different ASR models in an objective way.

Table 3: *Distribution of speech to be used in testing*

	Children	Teenagers	Elderly	Total
Female	19m	5m	17m	**41m**
Male	16m	6m	15m	**37m**
Total	**35m**	**11m**	**32m**	**1h18m**

Finally, because of the limited amount of data available for the Transitional region, it was not possible to train the language model of any of the ASR systems only on that speech. Therefore, the text used for training the language model contained the text of the entire JASMIN-CGN corpus, which allowed for building and training of the ASR models without issues afterward.

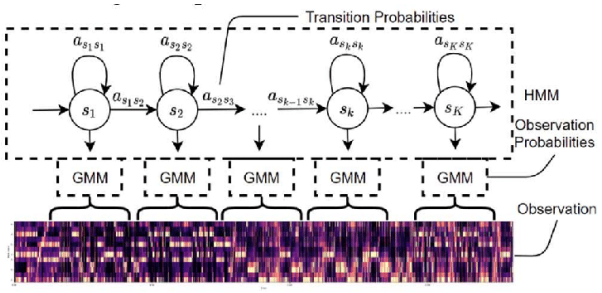

Figure 3: *GMM-HMM mixture model. Source: [17].*

3.2. Technical setup

3.2.1. Architecture of the ASR system

The architecture used for all ASR systems in this paper is based on a hybrid GMM-HMM acoustic model, as can be seen in figure 3, a trigram language model, and a dictionary/lexicon. The training is done using the Kaldi toolkit [15]. The code has been made publicly made available and a link to it can be found in the footnote[1] of this page. Due to the limited time span available for this research project and the low-resource setting of our data, this model is the only one being considered, but it is acknowledged that neural network models can also be used for such a task, and in most cases, they perform better than GMM-HMM models given enough data [8] [16].

3.2.2. SpecAugment Settings

A Python library has been used to achieve the augmentation, named Audiomentations [18]. The function used is *SpecFrequencyMask*. This function takes a mel frequency range as an argument which defines how much percentage out of the total audio file can the masked segment be. The minimum range, *min_mask_fraction*, that was used is 10%, and the maximum range, *max_mask_fraction*, is 20%. The function then takes an arbitrary number from that range as the range of the mask for an audio file, then applies masking randomly, which results in augmented data. There is also a parameter, p, of *SpecFrequencyMask*, which determines the probability that the mask will be applied to the audio file or not. In this case, since all of the data needs to be augmented, the parameter was set to 1 such that it is applied to the files with a 100% chance. The rest of the parameters remained with their default values.

3.3. ASR models trained

Four models have been trained and tested to evaluate SpecAugment's performance. The first model, *baseline*, contains only the original Transitional Dutch data from the JASMIN-CGN corpus, split as mentioned in the "Data preparation" section. The second one, *aug_sa*, has the original + augmented data using frequency masking from SpecAugment, with the settings described in the previous section. The amount of training data for this model is double in size compared to the original, due to the augmented data having the same duration as JASMIN-CGN's one.

In order to compare the results of SpecAugment with another augmentation technique that is more reliable for hybrid systems as well, the third model, *aug_vtlp* has been augmented

[1]Link to the files used throughout research: `https://github.com/greenw0lf/RP2022-JASMIN-ASR`

using VTLP, with the parameters and settings mentioned in the paper by Zhlebinkov [19]. Lastly, another model has been trained to evaluate the performance in correlation to the quantity of the data available, as the duration of the Transitional region from the JASMIN-CGN corpus amounts to only 6 hours and a half. Thus, the last model used in experiments, *tran_west*, contains speech from the corpus from both Transitional and West regions. The total amount of training data after adding the West region is 11 hours.

In order to understand the results of the last model mentioned better, a breakdown of the West region's distribution of speech is provided in table 4. The table has been obtained from the paper by Marinov [20], whose data quantity and augmentation work throughout the research project are similar to the author's.

Table 4: *Distribution of Western regional speech*

	Teenagers	Elderly	Total
Female	53m	1h46m	**2h39m**
Male	1h07m	45m	**1h52m**
Total	**2h**	**2h31m**	**4h31m**

There is no extra data for children provided, but there is quite a bit for teenagers and the elderly. It is expected that performance may be improved for those two age categories. When it comes to gender, there is more female speech than male, so a similar outcome is predicted.

3.4. Results

Results are measured using the Word Error Rate (WER) metric. WER is computed by dividing the number of errors, such as substitutions (words that were recognized incorrectly), deletions (words that the ASR did not detect and thus were deleted from the original text), and insertions (words that were detected when they should not have, and thus were inserted into the original text) by the number of words that were actually spoken.

Table 5 contains a breakdown of the results based on the different categories of speakers/speech that the questions cover.

Table 5: *WER performance (in %) of the different ASR models. Results highlighted in bold indicate the best results for each speaker/speech category*

	baseline	aug_sa	aug_vtlp	tran_west
Children	22.14	24.24	**21.79**	23.49
Teenagers	11.88	13.03	12.06	**11.65**
Elderly	19.43	20.19	19.63	**18.58**
Male	20.44	21.48	**19.82**	20.73
Female	**18.44**	20.1	18.79	**18.44**
Read	**9.46**	10.48	9.67	10.39
Conv/HMI	48.05	50.72	48.84	**46.61**
Overall	19.5	20.8	**19.44**	19.61

Looking at the baseline results, it can be seen that the ASR has significantly better performance for teenagers compared to children or the elderly. However, this could also be the case because of the amount of data available for this group compared to others (almost 4x less than either children or the elderly). Between children and the elderly, however, the difference is not as big, with elder speech doing better. The results are in line with the work of Feng et al. [4], who observed the best performance for teenagers, the second-best for people of age, and the worst for children.

When it comes to gender, however, the results are quite different. There is a 2% difference between male and female speech WER. This is comparable to the works of [4], [5] and [6], and contradicts the work of [7].

For read vs. conversational speech, the difference in performance is large. Read speech performs 5x better than conversational speech. This could be attributed to the difference in the amount of data between the 2 styles of speech since conversational data represents only 30% of the total duration of the speech. However, Feng et al. do not have such a stark difference between HMI (Human-Machine Interaction) and read speech, with the difference between them being only 13.7% [4]. This is an interesting finding that could be researched further with the entire data of the JASMIN-CGN corpus and this hybrid ASR architecture.

The model augmented with SpecAugment's frequency mask, aug_sa, did the poorest in performance compared to any other model. It does not manage to improve the WER whatsoever, instead, it worsens it and therefore it would not be indicated to use for this combination of ASR architecture + data.

The model augmented using VTLP (aug_vtlp), however, does the best overall in performance, although it is a tiny increase compared to the baseline (0.06%). It manages to achieve the best WER for children's speech and male speech. This proves that, by shifting children's vocal tracts, performance could be improved for this age group and the issue of shorter vocal tracts for this respective age group would be reduced.

Finally, the model that uses the training data of the Transitional region plus the entire data of the Western region, tran_west, is overall worse than the VTLP-augmented or the baseline models but does better in specific age/gender groups. The groups that benefit from the decrease in WER are teenagers, people of age, and female speakers. For teenagers and people of age, it does make sense because there has been more data for each of these 2 categories and there is no extra data available for children from the West region. It is also assumed that there is more varied speech and different speaking styles that add more robustness to the training of the model.

For female speech, there is an extra 47 minutes in the Western region compared to male speech, thus increasing the ratio of data between female and male speakers. It does not bring any improvement to the female group in terms of WER compared to the baseline, but it does worsen it for male speakers, thus making it more biased towards female speech.

The baseline model manages to achieve the best performance for *read speech*, whereas west_tran does that for *HMI speech*. When it comes to reducing the bias between the different age groups or genders or speech categories, aug_vtlp manages to reduce it the most, for both age and gender. By reducing the WER for children which was the worst-performing category in the baseline and increasing it for the other 2, the gap between the lowest and highest performance is reduced from 10.26% to 9.73%, and similar for gender, where the gap is reduced from 2% to 1.03%, almost half of the difference of the original.

The Transitional+West region model, in contrast, widens the gap, having the largest bias between the different groups. For age, the gap is 11.84%, and for gender, it is 2.29%. It does manage to reduce the gap between conversational and read speech, by 2.37%. This is interesting since a similar distribution to the Transitional region data has been observed in the West-

ern region as well. The reason could be that there was already too fewer data and, by adding slightly more, the performance difference can be considerable for this style of speech.

Going back to the SpecAugment model, it does manage to reduce the bias gap between male and female data by 0.62%, however, it does that by increasing the WER of both of them. However, for age, it widens the gap by almost 1%. The gap is also widened for read vs. conversational speech, by 1.65% compared to the baseline. Overall, it does not manage to reduce the bias compared to aug_vtlp or tran_west.

4. Discussion and conclusions

To conclude, SpecAugment fails to improve the performance of an ASR system built on the Transitional Dutch accent of the JASMIN-CGN corpus. The WER has not been lowered, on the contrary, it went up in all the different criteria of speakers/speech. When it comes to age, the difference between the individual groups is larger than that of the baseline, thus increasing the bias for age. The difference is quite significant in both cases, with the baseline having a 10.26% difference in performance between children and teenagers, 2.71% between children and the elderly, and 7.55% for teenagers and the elderly. The SpecAugment model, in turn, has 11.21%, 4.05%, and 7.16% WER. The performance obtained in both cases is similar to that of [4].

For gender, the differences are smaller. The baseline has a 2% difference in WER compared to SpecAugment which reduces that difference, by 0.62%. The difference in performance in both models is quite significant for the number of speakers each category has, but it is not as large as in the case of age. The results are in line with that of [4], [5] and [6].

Finally, for read vs. conversational, the baseline model achieves a 38.59% difference between the two, whereas the SpecAugment model has a gap of 40.24%. This difference, compared to the other two categories, is quite significant, and it has been majorly affected by the amount of HMI data versus read data (30/70 split of data).

The model that performed the best and reduced the bias for most criteria is the VTLP-augmented model. Therefore, it would be recommended to use this augmentation technique for this scenario in contrast to SpecAugment.

Possible improvements that can be done to this model are to add data from other regions as well, augment them using SpecAugment, and test the performance to see if the WER can be improved. The WER should be lower in that case and, if not, then it could be attributed to the combination of the augmentation technique + the architecture, which is the GMM-HMM hybrid model that was used throughout the project. Another idea would be to test the augmentation techniques implemented by my fellow researchers (SpecSwap, VTLP, pitch shifting, and frequency perturbation) plus SpecAugment on all of the JASMIN-CGN Dutch data (excluding Flemish) and see if an end-to-end model could be run that would reduce the bias for each accent simply by merging all of the data. The last idea that could be tackled is to analyze in-depth what the issue is with the SpecAugment model analyzed in this paper, if certain words or phonemes are hardly recognized or if some speakers are more prone to error, and do a comparison with the other models tested.

5. Responsible Research

Most of the applications involving ASRs are related to voice assistants and voice control and there are no apparent ethical issues that can be tackled by the author himself when it comes to this research. On the contrary, the author has attempted to reduce the bias of different genders or age groups by training and testing different ASR models. As for the research itself, the *Experimental Setup and Results* chapter should provide enough information for the results to be reproduced. The ASR models generated can contain different probabilities for the sequence of phonemes and words, which might lead to slightly different results when testing them, but the results obtained, if the experiment were to be reproduced, would be very similar, if not identical.

The JASMIN corpus that has been used as a dataset is a publicly available one with detailed documentation, developed by researchers from various Dutch research universities. Throughout this paper, it is clearly mentioned what the experimental procedure is and what speakers have been chosen from the corpus. There is no data that has been left out intentionally. Therefore, the author has managed to conduct research in a responsible and correct manner, without any external or internal biased influence.

6. Acknowledgments

I would like to thank my responsible professor, dr. Odette Scharenborg, for providing me and my research group with feedback for the final versions of the paper, as well as informing us whether we are on track or not.

Many thanks go to my research group, and especially to Alves Marinov, with whom I discussed together with the different augmentation techniques and gave opinions on each others' work, and to Nikolay Zhlebinkov, who has provided me with the tools necessary to augment my data using VTLP.

My greatest appreciation goes to my supervisor, dr. Tanvina Patel, which not only has helped me throughout the project with understanding the tools to be used and providing me with feedback throughout the stages of the research but doing so by rapidly replying to my requests and by providing considerable information.

Lastly, I thank the Delft High Performance Computing Centre [21] for providing me with access to the cluster where I conducted my experiments.

7. References

[1] N. Jaitly and G. E. Hinton, "Vocal Tract Length Perturbation (VTLP) improves speech recognition," 2013.

[2] 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.

[3] X. Song, Z. Wu, Y. Huang, D. Su, and H. Meng, "SpecSwap: A Simple Data Augmentation Method for End-to-End Speech Recognition," in *INTERSPEECH*, pp. 581–585, 2020.

[4] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying Bias in Automatic Speech Recognition," 2021.

[5] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?," pp. 2205–2208, 09 2005.

[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," 01 2013.

[7] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, (Valencia, Spain), pp. 53–59, Association for Computational Linguistics, Apr. 2017.

[8] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," 2018.

[9] 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," 2020.

[10] J. Poncelet and H. Van hamme, "Comparison of Self-Supervised Speech Pre-Training Methods on Flemish Dutch," 2021.

[11] D. Van Leeuwen, J. Kessens, E. Sanders, and H. van den Heuvel, "Results of the n-best 2008 dutch speech recognition evaluation," pp. 2571–2574, 09 2009.

[12] 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.

[13] H. Pérez-Espinosa, J. Martínez-Miranda, I. Espinosa-Curiel, J. Rodríguez-Jacobo, and H. Avila-George, "Using acoustic paralinguistic information to assess the interaction quality in speech-based systems for elderly users," *International Journal of Human-Computer Studies*, vol. 98, 2017.

[14] 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.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Vesel, "The Kaldi speech recognition toolkit," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 01 2011.

[16] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli, "Hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 312–317, 2013.

[17] T. Li, "Study on a CNN-HMM Approach for Audio-Based Musical Chord Recognition," *Journal of Physics: Conference Series*, vol. 1802, p. 032033, 03 2021.

[18] I. Jordal, "GitHub - iver56/audiomentations: A Python library for audio data." `https://github.com/iver56/audiomentations`, 2022. Accessed: May 30, 2022.

[19] N. Zhlebinkov, "Improving Northern Regional Dutch Speech Recognition by Adapting Perturbation-based Data Augmentation," 06 2022.

[20] A. Marinov, "Evaluating the Effect of SpecSwap for Purposes of Improving WER Performance of the Western Dutch Region Using the JASMIN-CGN Dataset," 06 2022.

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