Evaluating the Use of Pitch Shifting to Improve Automatic Speech Recognition Performance on Southern Dutch Accents

Amar Mešić
Supervisors: 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

Building Automatic Speech Recognizers (ASRs) has been a challenge in languages with insufficiently sized corpora or data sets. A further large issue in language corpora is biases against regionally accented speech and other speaker attributes. There are some techniques to improve ASR performance and reduce biases in these corpora, known as data augmentations. One audio data augmentation, pitch shifting, has had successes in other experiments for increasing ASR performance. Pitch shifting it is tested in this paper on the JASMIN-CGN speech data set from the Southern regions of the Netherlands. Using a hybrid GMM-HMM ASR, two baselines are developed, one using all speech data from the region, the other only using native speech. For the former ASR, pitch shifting is found to not improve Word Error Rate (WER) performance or reduce bias, but the latter succeeds in improving WER performance and reduced bias for certain speaker groups when augmented.

Index Terms: ASR, Data augmentation, Audio Augmentation, Speech Recognition, Pitch Shift, Hybrid ASR, Dutch, JASMIN-CGN, Bias

1 Introduction

In order to develop automatic speech recognizers (ASRs), computer scientists rely on large speech corpora. Humans are becoming ever more reliable on language technology, yet it remains difficult to produce training data. That is why researchers are looking into ways to more effectively make use of data in these corpora to make sure that more speaker groups can make use of these technologies.

The Dutch language has two main corpora: CGN (Corpus Gesproken Nederlands) and JASMIN-CGN. CGN was developed in 2004, containing nine million words as spoken in the Netherlands and Flanders [1]. However, CGN only contained data of spoken standard Dutch and only in two standard accents. Furthermore, CGN did not contain "speech of children, non-natives, elderly people and recordings of speech produced in human-machine interactions" [1].

Due to the limited representation of speaker groups, ASRs were not able to equally effectively recognize speech from all speakers when trained on CGN, such as regional accents. CGN was later extended, and JASMIN-CGN was formed, with the aim of including speech from previously excluded speakers groups. However, despite the improvements laid forward with the deployment of JASMIN-CGN, biases remain prevalent in ASRs trained on them. Feng et al. showed disparities in WER for regionally accented speech on an ASR trained on JASMIN-CGN [2], and other ASRs have also been found to perform differently for different regional accents [3, 4]. However, collecting even more data is a financially demanding and arduous task, which makes it impractical to continue including all possible speaker groups of a language in a corpus.

To tackle the deficit in data, data augmentation techniques have been shown to effectively improve performance [5, 6]. In this paper, research will be conducted into the pitch shifting data augmentation technique, which will be evaluated on ASRs trained and developed on Southern Dutch speech from JASMIN-CGN. Pitch shifting has been found to be an effective audio data augmentation method for improving ASR performance and with singing voice recognition [7–10]. Spectral characteristics of users have also been found to make certain speakers more intelligible than others, showing the potential of an augmentation that alters spectral characteristics [11, 12].

The aim of this experiment is to answer the following research question: Can augmenting data from the existing JASMIN-CGN corpus using pitch shifting improve ASR performance on southern Dutch accents? This question can be further divided into three subquestions:

- Is it possible to get an improved word error rate (WER) on an ASR trained with augmented Southern Dutch data from JASMIN-CGN?
- Is it possible to get an improved WER for children and the elderly on an ASR trained with augmented Southern Dutch data from JASMIN-CGN?
- Is it possible to get an improved WER for non-native speakers on an ASR trained with aug-

mented Southern Dutch data from JASMIN-CGN?

- Is it possible to reduce the difference in WER for male and female speakers on an ASR trained with augmented Southern Dutch data from JASMIN-CGN?

The next section formulates the methodology used to approach answering the research question and introduces pitch shifting. Section 3 lays out the tools as well as the set-up used in the experiment. Section 4 unpacks the results of running pitch shifting, which are discussed in Section 5, and the research questions are answered in Section 6. Section 7 also touches on ethical implications and how research was carried out responsibly.

2 Methodology

We take a deep look into the structure of the JASMIN corpus, which contains speech data from all dialect regions from various speaker groups. Pitch shifting, the data augmentation used, and its use cases are presented along with brief overviews of other similar augmentation techniques.

2.1 JASMIN-CGN Corpus

The JASMIN-CGN corpus, contains 90+ hours of spoken Dutch collected in different regions by both natives and non-natives, in different age groups, in different speech environments. The corpus contains over 500 speakers, with 154 speakers from the Southern Dutch region. JASMIN contains five age groups:

1. *children (7-11)*
2. *native teenagers (12-17)*
3. *non-native teenagers (12-18)*
4. *adults (19-65)*
5. *elderly (65+)*

Natives are categorized in groups (1, 2, 5), while non-natives are in groups (3, 4). The language proficiency of Dutch for all non-native adults ranges from A1 to B2.

Furthermore, JASMIN includes a substantial amount of regional speech in order to reduce regional biases, namely from the Northern, Transitional, Western, Southern, and Flemish regions. The speech itself is split into two components, conversational and read. The former intends to simulate human-machine interaction, an ever growing field of computer science.

This paper is mainly concerned with speech from the Southern Dutch dialect regions, which contains speech from North Brabant. There is 22.53 hours of speech from this region, and table 2 shows the time divided by different speaker groups.

2.2 Data Augmentations

2.2.1 Pitch Shift

Pitch shifting stretches or withdraws the spectral profile of speech, while maintaining the tempo of the audio. Pitch shifting is able to make listeners perceive speech as coming from another speaker [13]. Pitch is different to frequency due to the way humans perceive the frequency of noise. The Mel scale measures the human perception of frequency i.e., pitch compared to the real measured frequency of noise [14]. A function that well approximates the relationship between frequency and pitch is a skewed logistical function: $F_M = \frac{f}{af+b}$ where F_M is the Mel frequency (i.e. pitch), and f being the frequency.

In the Prodorshok I, a small-scale isolated data set for Bengali, pitch shifting was used to improve ASR performance for a Human-Computer interaction scenario with a relative improvement of 12.4% [7]. The data set used in the study, however, consists of only 30 utterances from 35 speakers. Furthermore, the study misses out on conversational HMI speech, which will be considered with JASMIN. Another study from Schlüter & Grill managed to achieve a relative improvement in classification error of 21.6% when applying the pitch shift augmentation to their data sets in a task of singing voice detection [10]. Pitch shifting was done within a range of $\pm 50\%$ with performance varying slightly for values within that range. A pitch shift of -40% can be seen in figure 1. They also found that pitch shift could be improved a further 6% by augmenting the data further. Augmenting the test set was also found to further improve performance. This study shows promising results on a CNN, whereas this paper will focus on a GMM-HMM model.

The way pitch shifting was implemented by [10] was by representing the data through a mel spectro-

gram, which is scaled while retaining an anchor at 0Hz. A lot of preprocessing steps do not need to be taken in our case due to preprocessing that is automatically done by Kaldi (discussed in Section 3.1).

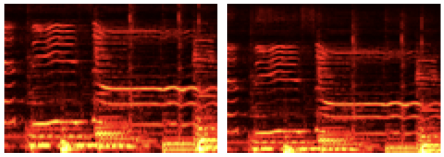

Figure 1: *A spectrogram, with the frequency on the y-axis and time on the x-axis, shows the intensity of the different frequencies over time for a speech segment. The left image shows the spectrogram of a sample segment, while the right image shows the same segments augmented with pitch shift of -40%.*

Another study found that cochlear implant were able to improve speech recognition with speech that had pitch shifting and spectral normalization applied [15], but only takes gender into consideration when looking at recognition improvements. We will also look at how age and nativity impacts performance improvements.

3 Experimental Setup

3.1 ASR set-up and Kaldi

Kaldi is an ASR toolkit [16] to develop and train an ASR. The type ASR system used for this experiment consists of a hybrid GMM-HMM acoustic model and a trigram language model. JASMIN contains a lexicon which allows an ASR to train on the transcribed speech. Due to time constraints for this project this is the only model considered, leaving the possibility to extend this research on a DNN-based model.

3.2 Audiomentations

Audiomentations [17] is an open-source library that provides audio editing and augmentation functionality. Pitch shift is a provided augmentation, and will be thus applied on the data. It takes in three parameters: minimum semitones ($min\_semitones$), maximum semitones ($max\_semitones$), and p, the probability (0,1) of applying the augmentation. The probability value should be set to 1 by default if applying pitch shift to all segments. In our case where we seek to shift the pitch equally for all speakers, we set $min\_semitones = max\_semitones$. The use of semitones ensures that we are not just shifting frequency, but specifically pitch.

3.3 Semitones

Audiomentations uses semitones as the unit for pitch, so a percentual change in shift should be translated into a semitone shift. An increase/decrease in a semitone results in a uniform increase in pitch. There are twelve semitones distributed within an octave, and octaves are an interval where the higher note has double the frequency of the lower. The semitone thus has the ratio:

$$r^{12} = 2$$

$$r = \sqrt[12]{2} \approx 1.05946$$

The ratio can then be used to obtain the number of semitones S required for a shift of a percentage P.

$$\log_r (1 \pm \frac{P\%}{100\%}) = S$$

3.4 Baseline WER Performance

Using the JASMIN documentation [1], the Southern Dutch speech data was singled out from JASMIN, of which an 80% - 20% split was made. It was ensured that the train and test set had a disjoint set of speakers, and were proportionally split among gender, age group, and nativity. Table 2 shows that the duration of the speech i.e., the size of the train and test set were split equally (80% - 20%) among all facets. The table also shows that females have more speaking time than males, with the split being 58% - 42%, and similarly non-natives are represented more than natives, with a split of 73% - 27%.

The Kaldi [16] GMM-HMM hybrid ASR model trained on the data resulted in a 43.48% WER. Additionally, another natively trained (NT) baseline model was set up, with this second one only trained on native speakers. This model thus only has 4.89 hours of training data. The reason a second baseline with only native speakers has been made is due to the fact that native speakers are much more likely to possess the regional Southern Dutch accent, and

WER (%)	Baseline	Pitch Shift					Natively Trained (NT)	
		$+30\%$	-30%	$\pm30\%$	$+50\%$	-50%	Baseline	$\pm30\%$
Combined	43.48	44.65	54.41	46.14	45.46	54.19	60.02	45.86
Conversational	62.53	63.48	72.34	64.71	63.27	73.85	74.53	64.78
Read	37.1	38.23	46.63	39.93	38.81	47.34	54.97	39.62
Male	43.37	44.84	55.64	46.2	45.3	56.01	60.19	46.22
Female	43.29	44.31	53.32	45.85	44.94	52.76	59.83	45.42
Age Group 1	52.24	53.69	65.39	53.61	53.35	64.1	55.4	53.66
Age Group 2	19.17	21.63	36.49	27	22.69	32.61	14.6	25.43
Age Group 3	42.38	43	52.71	44.48	43.9	51.98	65.62	44.35
Age Group 4	41.18	41.5	50.39	42.45	41.9	52.51	72.69	41.87
Age Group 5	55.91	57.54	64.89	59.67	57.74	66.1	54.39	58.9
Native	44.47	46.38	57.18	48.41	46.47	55.98	43.47	48.02
Non-native	42.06	42.61	51.75	43.72	43.29	52.44	68.91	43.67

Table 1: Results table containing the WERs for different speaker groups under different augmentations. A lower result, corresponding with a lighter background, indicates better performance.

Table 2: The amount of speaking time of Southern Dutch for different speech types and speaker groups in hours.

(hours)	Train	Test	Total
Conversational	5.02	1.0	6.02
Read	13.45	3.07	16.52
Male	7.89	1.65	9.54
Female	10.57	2.41	12.98
Native	4.89	1.26	6.15
Non-native	13.57	2.80	16.37
Age Group 1	1.58	0.41	1.99
Age Group 2	1.33	0.32	1.65
Age Group 3	5.77	1.43	7.2
Age Group 4	7.71	1.38	9.09
Age Group 5	2.08	0.52	2.6
Total time:	**18.46**	**4.06**	**22.53**

improvements from the second baseline would indicate equally strongly that pitch shifting is an effective data augmentation. The WER for the native baseline is significantly higher at 60.02%.

3.5 Augmentations Applied

The augmentations that will be applied on the training set will be:

$$-30\%, +30\%, \pm30\%, -50\%, +50\%$$

Table 3: The amount of speaking time of Southern Dutch after augmenting, in hours.

(hours)	Train	Test
Baseline	18.46	4.06
PS +30% PS -30% PS +50% PS -50%	36.92	4.06
PS $\pm30\%$	55.38	4.06

This was a set of percentages which provided the greatest improvement in [10]. Also, an augmentation of $\pm30\%$ will be applied to the NT baseline. All augmentations prefixed with a $'-'$ represent a training set with the original data and the data shifted down by the percentage, and shifted up for $'+'$. The percentage must be converted to a semitone value, so for example $+50\% \rightarrow \log_{1.05946}(1 + \frac{50\%}{100\%}) \approx \log_{1.0595} 1.5 \approx 7.02$. A $+50\%$ pitch shift would thus be shifted by 7.02 semitones, and would result in the training set doubling in the amount of hours. For $\pm30\%$, we pitch-shift the data both up *and* down, resulting in a tripling of the train set size. the amount of training data after augmenting is visible in table 3.

4 Results

Multiple setups with different variations of the pitch shift augmentation were run, providing WERs both for the combined test set, which represent the all the speakers in the test set, as well as for the separated subgroups, which separates the tests set into disjoint subsets based on speaker characteristics. These subgroups were separated on age, gender, nativity, and the type of speech (Conversational/HMI or Read).

The results are measured using Word Error Rate (WER), which is attained by dividing the number of errors for a word by the number of words actually spoken. An error can be either a substitution (wrong word), insertion (word detected where there is not one), or deletion (word detected not when there is one).

For the WER of the combined/total test set, the original data provides the lowest WER, meaning no augmentation was able to lower the WER for Southern Dutch data. Excluding the baseline, the best performing data augmentations were $+30\%$, $\pm 30\%$ and $+50\%$. The worst performing being -30% and -50%, with the only difference between the good and bad augmentations being the direction in the pitch changes. It is interesting to note that $\pm 30\%$ does not perform better than $+30\%$, since it provides additional training data.

However, when training only on native speakers, the data augmentation *is* able to reduce the WER with a relative improvement of 23.6%. The results for the NT ASR are discussed in-depth in section 4.1.

Looking at the speech components, conversational/HMI speech deteriorated the least with the $+50\%$ pitch shift with a relative rate of 1.2% but the most with the -50%. For read speech, deterioration was minimal with the $+30\%$ with a relative rate of 3.0% shift and maximal with -50%.

When it comes to gender, throughout all augmentation cases, female speakers consistently have lower WERs than their male counterparts. However, the difference is not significant, as the largest percentual improvement was 4.3%, and this was on the same augmentation that provided the worst WER for both test sets.

Moving on to the different age groups, once again all age groups suffer from minor deteriorations in the WER. In consistency with the combined test set and that for both genders, downward pitch shifts are generally harmful, while upwards shifts remain close to their original values. The age group with the smallest deterioration was group 4, which consists of non-native adults, with a relative deterioration of 0.8%. This was achieved with a $+30\%$ pitch shift.

When looking at speakers categorized by nativity, again there is no improvements, but we can see the lowest and largest deterioration. The lowest for natives is a relative deterioration of 4.3%, and 1.3% for non-natives, both with a $+30\%$ shift. The worst augmentations for natives and non-natives are -30% and -50% respectively.

4.1 Improvements when Natively Trained

When looking at the last two columns of table 1, there does seem to be some improvement in the WER for certain rows. Firstly, the WER for the combined speaker test set shows a significant relative improvement of 23.6%. This is reflected in both speech components, having a relative improvement of of 13.1% for conversational speech, and an improvement of 27.9% for read speech. Female and male speech recognition had a relative improvement of 23.2% and 24.1% respectively. With no significant difference between the two, both can be said to have improved equally well. In the age groups, groups $1, 3, 4$ all had improvements, while $2, 5$ got significantly worse. Lastly, looking at performance for the native speakers on the NT ASR, performance deteriorated by 10.5%, while it improved for non-native speakers by a drastic 36.6% relative WER improvement.

5 Discussion

The results come with two different conclusions for the two different baselines, showing that using pitch shift to augment the complete training data does not lead to improvements in ASR performance, but using it on native-only training data does improve performance as well as reduce bias.

When it comes to the fully-trained baseline, none of the pitch shift strategies improved baseline performance, and furthermore they did not improve performance on any of the test subgroups. As mentioned, the best performing pitch shift strategies were $+30\%$, $+50\%$, and $\pm 30\%$ in order from best to worst. There is no significant difference for

$+30\%$ and $+50\%$ in the baseline WERs, with an absolute difference of 0.81% or a relative difference of $1.8\% < 5\%$. Additionally, a pitch shift of $\pm 30\%$ performed 3.3% worse relatively despite having more data.

The reason for positive or upward pitch shifts being more effective could be due to the fact that teenagers are found to be the most intelligible speaker age group by far in all experiments as can be seen in table 1. While there may be multiple factors for making teenagers more intelligible, they certainly have higher-pitched voices than adults [11]. Therefore an upward pitch shift on adult speech makes the spectral profile more similar to that of adolescents, and could be a key factor in the better performance for $+30\%$ and $+50\%$ pitch shifts.

The cause for a worse performance for $\pm 30\%$ has to be because the downward shifted audio is causing the model to fit to unrealistic audio data. The hypothesis for $\pm 30\%$ was that speakers with a high-pitched voice would provide useful training data for their downward shifted audio, and likewise for those with low-pitched voices. This would then have created more audio data closer to the mean speaker pitch, but what $\pm 30\%$ would not do is reduce the variance of pitch. Another approach where $\pm 30\%$ is applied selectively only to speakers with low- or high-pitched voices could reduce the pitch variance, possibly making a more accurate ASR. On the other hand, the worst performing pitch shift strategies were -30%, -50%, indicating that downward shifts should be avoided when pitch shifting.

We also see that females have slightly better WERs than their male counterparts, confirming [2] and [12]. The same performance difference goes for non-native speakers as well as for read speech. In all three of these cases, the group with more hours of training data has lower WERs. This indicates that there is a positive correlation between representation and performance, and can be investigated further by measuring performance differences if equal training times were given.

On the topic of bias, compared to the full baseline, no augmentation has managed to reduce bias for any speaker group. On the other hand, for the NT ASR, a bias was reduced against non-native speakers. For the NT baseline there was an absolute difference of 25.4% between native and non-native WERs, but the difference dropped to 4.4% after applying $\pm 30\%$. Furthermore, if looking only at native speakers (age groups $1, 2, 5$), there is also a decrease in bias against native children for the NT ASR. The absolute difference between child WER and native WER is initially 11.9%, but drops to 5.6%.

The NT ASR does provide improvements in WER when pitch shift is applied, but interestingly, it does not improve performance on native speakers, and while pitch shifting the NT ASR reduces bias against non-natives, it is still more biased than the normal baseline. What remains to be seen then is how performance would improve in other dialect/accent regions in JASMIN, since the other three regions contain $> 95\%$ native speakers.

6 Conclusion and Future Work

All in all, on an ASR trained on Southern Dutch speech does not improve WER performance when pitch shifting is applied on training data, and does not reduce bias. The large presence of non-native speakers and the variety in their accents is likely the reason it is so difficult to improve an ASR that is trained on it. Upward pitch shifts were found to be more effective in not deteriorating performance, and combining upward and downward pitch shifts was not found to be more effective. On a natively trained ASR, performance is improved and bias is reduced for native children and non-native speakers. Due to the lack of a significant difference in male and female WERs in both baselines, there was no gender bias present. Further experiments can be done on the Southern Dutch speech data with different data augmentations such as vocal tract length perturbation and frequency perturbation. Testing pitch shifting on other dialect/accent regions as well as on the JASMIN-CGN corpus as a whole could provide more insights on the effectiveness of pitch shifting to improve WER performance and reduce bias.

7 Responsible Research

It is important to consider the ethical implications of this study. The data dealt with and experiments conducted have few ethical implications, but the results have a broad impact, as this study shows how bias can be mitigated in data sets. The experiment should also be reproducible.

The data used in this study is courtesy of the JASMIN-CGN data set, which is part of the public domain, and was developed by researchers working under the oversight of Dutch and Belgian governments under the CGN-bureau [18]. This includes the recruitment of all speakers as well, so ethical considerations of confidentiality are not an issue, and the augmented data is label-preserving.

This study attempted to reduce bias in speakers groups, and it showed that, given a data set of native speakers, it is possible to reduce bias against certain speaker groups, which has positive ethical implications.

The steps taken to augment the data have also been described in Section 3, allowing others to recreate the results obtained in this experiment. Any differences caused may occur due to differences in the train/test split, which should be negligible. The technologies and tools used are all open-source, and have been fully credited.

8 Acknowledgements

I would like to express my gratitude to my supervisor for this project, Tanvina Patel, who was eager to help while we were accustoming ourselves to the tools we had to use, and was always readily available to run our jobs. I would also like to thank my responsible professor for always cheerfully guiding us in the right direction, and my colleagues, who were willing to share their tools and findings to help me avoid any pitfalls. Lastly, I would like to acknowledge my use of the Delft Blue supercomputer provided to me by the Delft High Performance Computing Centre [19].

References

[1] Catia Cucchiarini, Hugo Van hamme, Olga van Herwijnen, and Felix 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, May 2006. European Language Resources Association (ELRA).

[2] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. Quantifying bias in automatic speech recognition, 2021.

[3] Majdi 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. 2013.

[4] Rachael Tatman and Conner Kasten. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *INTERSPEECH*, 2017.

[5] Shakti P. Rath Anton Ragni, Kate M. Knill and Mark J. F. Gales. Data augmentation for low resource languages. *Interspeech*, 2014.

[6] Matthew Baas and Herman Kamper. Voice conversion can improve asr in very low-resource settings, 2021.

[7] Mohi Reza, Warida Rashid, and Moin Mostakim. Prodorshok i: A bengali isolated speech dataset for voice-based assistive technologies: A comparative analysis of the effects of data augmentation on hmm-gmm and dnn classifiers. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 396–399, 2017.

[8] Naoyuki Kanda, Ryu Takeda, and Yasunari Obuchi. Elastic spectral distortion for low resource speech recognition with deep neural networks. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 309–314, 2013.

[9] Ishwar Chandra Yadav and Gayadhar Pradhan. Significance of pitch-based spectral normalization for children's speech recognition. *IEEE Signal Processing Letters*, 26(12):1822–1826, 2019.

[10] Jan Schlüter and Thomas Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR*, 2015.

[11] Elaine Stathopoulos, Jessica Huber, and Joan Sussman. Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4-93 years of age. *Journal of speech, language, and hearing research : JSLHR*, 54:1011–21, 08 2011.

[12] Ann R. Bradlow, Gina M. Torretta, and David B. Pisoni. Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20(3):255–272, 1996. Acoustic Echo Control and Speech Enhancement Techniques.

[13] Chuping Liu, John Galvin, Qian-Jie Fu, and Shrikanth S. Narayanan. Effect of spectral normalization on different talker speech recognition by cochlear implant users. *The Journal of the Acoustical Society of America*, 123(5):2836–2847, 2008.

[14] S. Umesh, L. Cohen, and D. Nelson. Fitting the mel scale. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, volume 1, pages 217–220 vol.1, 1999.

[15] Chuping Liu, J. Galvin, Qian-Jie Fu, and Shrikanth Narayanan. Effect of spectral normalization on different talker speech recognition by cochlear implant users. *The Journal of the Acoustical Society of America*, 123:2836–47, 06 2008.

[16] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Vesel. The kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 01 2011.

[17] iver56. audiomentations. https://github.com/iver56/audiomentations, 2022.

[18] Het project corpus gesproken nederlands. Available at https://lands.let.ru.nl/cgn/doc_Dutch/topics/project/pro_info.htm#intro.

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