Improving ASR performance on Jasmin Flemish Dutch data by performing frequency perturbation

Neal Sweijen
Supervisor(s): Odette Scharenborg, Tanvina Patel
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

ASR (automatic speech recognition) systems are used widely in our current day and age. However, for a technology that is used so much in our daily life it contains a lot of bias. This means that not all people can use it equally, people with a different gender, age and dialect will all see different results. The goal of this paper is to reduce this bias, in this case the dialect Flemish Dutch by increasing the performance of this dialect. Since collecting data is expensive, a data augmentation technique has been used. This technique has been used to increase the training data and lower the word error rate of this dialect. Frequency perturbation was used as the data augmentation technique. This technique amplifies or reduces the amplitude of certain frequency bands. We managed to improve upon the Flemish Dutch dialect slightly. Even though the dialect is still quite a bit worse compared to other Dutch dialects, it was improved nonetheless.

1 Introduction

Automatic Speech Recognition systems are widely used in this day and age, they are used in a wide array of devices such as Siri and Google Home, we use them to document important notes and much more. It becomes clear that everyone has the potential to benefit from such systems. However, unlike other technologies that require more traditional input, like text or images, ASR is a technology that can differ heavily in performance based on how different people speak. Bias however, is of course something undesirable, everyone should be able to use ASR systems and benefit from its uses. This is why it is important to improve upon the performance of ASR's by increasing the recognition rate, in this case of Flemish Dutch. Flemish dutch especially, is a bad performing dialect when it comes to ASR [1]. So it is invaluable to improve its performance.

Other work related to this one has already been done, for example for accented vs standard British English speech [2]. This research investigated how to compensate the effects of regional accents on ASR systems by using different acoustic modelling techniques. In another research it was found that there are a lot of biases in ASR based on different factors such as gender, age and region of the speaker in the Arabic language [3]. They found that these different characteristics can influence performance. Females for example had better results compared to male speakers. Data augmentation is another technique used for consistent performance gains in ASR systems [4]. In this research it was shown that two of the worst performing languages in a speech corpus called the Babel corpus, Assamese and Zulu, had consistent performance gain. This research used two forms of data augmentation namely: vocal tract length perturbation (VTLP) and semi supervised training. This shows us that, even if there is no current work on this specific dialect (Flemish Dutch), it seems very possible to increase performance for an ASR system with data from an existing corpus.

Since collecting data for a Corpus is quite an expensive task in this research we opted for a data augmentation technique. Data augmentation is often used to improve the quality of databases [5]. Data Augmentation techniques are used to modify the existing data, in this case the Jasmin-CGN corpus [6], in order to increase the amount of training data you have. This increase in data is used to increase performance in an ASR system. Because of our low budget we opted for data augmentation techniques that could be easily implemented. There has been quite some work in this field and there are a lot of interesting augmentation techniques we can choose from. For example speed perturbation and VTLP look promising [5] [7]. Since it has been shown that they can improve various ASR systems. There are also data augmentation techniques that act on the frequency spectrum of an audio-file. Specaugment for example is a technique where the frequency domain is used to mask various frequency and time bands and seems to be very promising when considering speech data since it has also shown to increase performance [8]. What these techniques have in common is that they all increase the speaker variety. This is because when there are used on an audio-file they alter the sound in some way causing the creation of 'new' speech. In this research, frequency perturbation was used. Since there is no recent work to be found on frequency perturbation but there seems to be some potential, this is the technique that was chosen. The potential comes from the fact that frequency perturbation can also cause more speaker variety since it can too alter an existing audio-file and thus might increase performance.

This paper aims to answer the following research question: Can data augmentation improve the ASR performance on Jasmin Flemish Dutch data? To help answer this question, three additional sub-questions will be used: Can we get a WER (word error rate) better than the original baseline for the three speaker groups of children, adults and older adults. These three groups were chosen to see if there will be a difference in performance in the individual groups. Because our data augmentation technique might cause one group to perform better than the others. These three groups do a good job representing the whole of the corpus. Furthermore, there will be additional speaker groups apart from the three mentioned above that will be tracked. This will be done to find some unexpected causes for our performance to not do as expected.

In the next section the methodology will be handled, here the experiment that will be used to answer the research question as well as the different methods used will be explained. After that, in section 3 the experimental setup will be described in detail together with the results of this experiment. Section 4 will summarize the results of section 3 and discuss any improvements and questions for the future. Finally in section 5 the ethical aspects that affect this research will be handled.

2 Methodology

Now our method will be defined in order to answer the research question. First, the evaluation of different ASR systems is handled. After which it is important to understand the data used: the Jasmin-CGN corpus. Next to define the speech recognition toolkit used. Lastly it is beneficial to go over the data augmentation technique used: frequency perturbation.

2.1 Evaluating ASR systems

In this research different ASR systems will be trained. Starting with a baseline system that will be created from the Jasmin-CGN corpus. This ASR system will be trained using various tools including the Kaldi toolkit. This system will be evaluated by looking at the WER of various speaker groups and type of speech. The baseline system will be used to compare an ASR without augmentation with systems that do use augmented data, in order to see if we can increase the recognition rate. After this, frequency perturbation will be applied on the audio-files of the Jasmin-CGN corpus after which these newly augmented files will be used as well as the original files to train a new ASR system. This new augmented system will be evaluated again. This process will be repeated a couple of times to find the most improved ASR system. Using this method it is possible to find whether data augmentation can improve the performance of Jasmin Flemish Dutch data.

2.2 Jasmin-CGN corpus

The Jasmin-CGN [6] is an extension of CGN corpus (corpus gesproken Nederlands) [9]. The Jasmin corpus will used to create the baseline ASR system and the consequent data augmented ASR systems. The data from this corpus is, contrary to the CGN corpus, annotated by speaker groups, gender, nativeness, age, mother-tongue, proficiency in Dutch, region and dialect. Aside from these variables, the corpus is also divided in read and conversational speech. Both are combined to create the ASR systems in this paper. Additionally, this paper is only interested in the Flemish region which consists of 4 dialects: West-Flemish (peripheral region), East-Flemish (transitional region), Brabant (core region) and Limburg (peripheral region).

In table 1 you can see how the groups important for this paper are distributed. As said before in the introduction for this research the sub-questions focus on children, adults and older adults. Since the children are split into three groups, they are evaluated separately in order to get the most detailed results. Something else to notice is that because they consist of three groups, there are much more children speakers compared to adults and older adults. Another aspect of the Jasmin corpus is that it includes non-native people as well. Even though the level of speech of these non-native speakers is substantially lower than the native speakers, there are taken into account. If the decision was made to leave them out of this research, the data set would have been much smaller and they would have missed the core speaker group of adults.

2.3 Kaldi

The actual ASR systems are trained using various techniques one of them is the kaldi toolkit [10]. Kaldi is a toolkit for speech recognition intended for use by speech recognition researches. Kaldi is perfect for building recognition systems.

2.4 Frequency perturbation

As said before, frequency perturbation will be implemented and used to create the ASR systems. The data augmentation technique will be used on the original data, so that we get additional data. Then this additional data combined with the original data are used to create a new ASR system. This section will go over the chosen data augmentation technique.

Speaker group	Number of speakers
Native children ages 7-11	43
Native children ages 12-16	44
Non native children	52
Non native Adults	30
Native Adults adults above 65	38
Male	96
Female	111

Table 1: Speaker distribution in Jasmin corpus for Flemish Dutch

Frequency perturbation is quite a simple technique. What it does is simply amplify or reduce the volume of certain frequency bands of the data randomly. In this modified implementation from audiomentations [11] there are 7 frequency bands: 42-95hz, 91-204hz, 196-244hz, 421-948hz, 909-2045hz, 1957-4404hz and 4216-9486hz. Each of these frequency bands will get boosted or cut by a random amount of decibel. This decibel range is specified in the script for example: -12db-12db. Which will cause a frequency band to be cut by maximum of 12 and boosted by at most 12 decibel.

In this implementation three different kinds of filters are used to get this effect: First a low shelf filter is applied, followed by five peaking filters and finally a high shelf filter is applied. A low shelf filters allows you to cut or boost the low end of the frequency spectrum, a high shelf filter allows this for the high end of the frequency spectrum. The peaking filters do this for all the frequency bands in between the low and the high end.

3 Experimental setup and results

In order to answer the aforementioned posed research question, it is necessary to perform an experiment. In this section it is explained how each ASR system was created. This will be done by first going over how the train/test data was setup and secondly by explaining how the training was actually performed. Finally, the results of this experiment will be handled.

3.1 Experimental setup

Baseline ASR

First, all the Flemish data of the Jasmin-CGN corpus was prepared. This gave files with all info necessary to create an ASR of this data. The files necessary were text, utt2spk, wav.scp and segments. From each of these files a train/test split was created. In this experiment a 90/10 split was used. It was of utmost importance that not only the data was just split using 90 and 10 percent of the data but the characteristics of the speakers, mentioned in Jasmin-Cgn corpus subsection were also distributed 90/10. This would make sure that the train and test data contained a representative amount of each gender, region etc. Because this split, bias towards a single speaker group was attempted to be avoided. Additionally speakers are not allowed to overlap in these two sets. In table 2 you can see the train/test split used. From this training data a baseline ASR was created using Kaldi. Then the test data was used to obtain a WER from this system.

Number of speakers	test	train
Native children ages 7-11	4	39
Native children ages 12-16	5	39
Non native children	6	46
Non native Adults	2	28
Native Adults adults above 65	4	34
Male	9	90
Female	12	99

Table 2: Train/test distribution

Augmented ASR

After having created the baseline ASR, another augmented ASR was created. First all the .wav files used in the baseline were copied locally. On each file frequency perturbation was applied using a python script made for this experiment. Now with the data augmented this extra data was put in the corresponding folders. After this, the original train files were expanded to, besides just containing the original data, now also containing the augmented data. Using this new training data a new ASR system was created. Then again the same test data as for the baseline ASR was used to obtain a WER from this new system. This process was done three times in total, to create three augmented ASR's. For each a different minimum and maximum gain dB were used. For ASR system Aug1 the parameters -12 minimum gain and 12 maximum gain was used. For ASR Aug2 -24 minimum gain and 12 maximum gain was used. For ASR Aug3 -36 minimum gain and 36 maximum gain was used.

Combined ASR

Finally one last ASR system was created. This last ASR, AugC, contained the training data from the baseline and the augmented data from each of the three augmented ASR's: Aug1, Aug2 and Aug3. Again, the same test data as all the other ASR systems was used. In table 3 it is shown how much hours of train and test data each of the ASR systems have.

	Baseline	Aug1	Aug2	Aug3	AugC
Train	22.13	44.28	44.28	44.28	87.91
Test	2.79	2.79	2.79	2.79	2.79

Table 3: Distribution of training/testing of each ASR system in hours

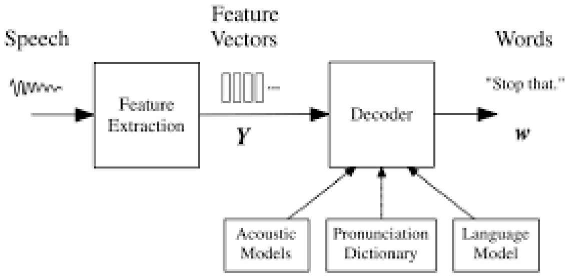

Figure 1: ASR system from: [12]

3.2 ASR architecture

The training of all the ASR systems has been done the same. As mentioned in the methodology the Kaldi toolkit has been used to train the acoustic model. For this a Gaussian Mixture Model/Hidden Markov Model or GMM/HMM is used. Apart from that the lexicon from the CGN corpus has been used. The language model is created from the training data. All of this together creates the ASR systems. All this was done by executing a single script so it is easily reproducible. In figure 1 a schematic of training an ASR system is shown.

3.3 Results

The performance as mentioned in the introduction is measured using the WER rate. Which is calculated as the total number of insertions, substitutions and deletions divided by the total number of words. In tables 4 and 5 the results of the baseline ASR, the three ASR's with different augmented data and the combined ASR are shown. There are 10 WER results per ASR system. Combined is the total WER of all the test data. Native children ages 7-11, native children ages 12-16, non native children, non native adults, native adults above 65, male and female all show the WER of the respective speaker group in the test data. Read contains the WER of read speech part of the test data and conversational contains the WER for the conversational part of the data data.

	Baseline	Aug1	Aug2
Combined	47.14	46.01	45.03
Native children ages 7-11	51.22	48.87	47.08
Native children ages 12-16	37.91	37.15	36.29
Non native children	54.16	52.97	52.27
Non native Adults	56.21	55.20	53.91
Native Adults adults above 65	45.90	45.92	45.10
Male	54.22	52.17	51.23
Female	45.5	42.93	41.93
Read	34.88	33.87	33.15
Conversational	72.51	72.13	71.03

Table 4: Percentage word error rate scores for each ASR system

Results Baseline

The results of the baseline are on par with the results from the earlier mentioned research [1]. Which is a research that quantified biases in Automatic speech recognition, one of them being Flemish. There are some differences, one of them being that the conversational speech and older adults speaker group are performing a bit worse. But the the speaker groups teens and children are performing quite a bit better in this baseline. These changes can be attributed to number of differences between to the two systems. For starters, in this baseline more read speech than conversational is used for exam-

	Aug3	AugC
Combined	44.65	45.59
Native children ages 7-11	46.75	48.12
Native children ages 12-16	36.34	37.41
Non native children	51.82	51.82
Non native Adults	53.86	54.49
Native Adults adults above 65	44.71	45.58
Male	50.89	51.56
Female	41.60	42.62
Read	32.09	32.76
Conversational	72.61	73.74

Table 5: Continuation of table 4: percentage word error rate scores for each ASR system

ple, this is because the Flemish dialect in the Jasmin corpus contains quite a bit more read speech compared to conversational. Other than that in our ASR system non-native speakers are also used for training, which was not the case in the aforementioned research. Another major difference is that in this experiment, the training data only consisted of the Flemish region, in the other research it was trained without a specific regional accent. Also the Flemish data in the Jasmin corpus has a lot of children speakers in comparison, which could result in the better performing children groups. The last difference is that in the other research a DNN-HMM (deep neural network) system was used as opposed to this research which uses GMM-HMM. DNN-HMM have been known for better performance [2]. We can also see that females are performing a lot better than the male speakers with a ten percent gap between the two. While female speakers also have a lower WER in the previous mentioned research, the gap between male and female is not as high.

Results Augmented ASR

When looking at the three augmented ASR systems, it is clear to see that every system improved the WER a small bit compared to the previous one. Aug1 performs better than the baseline, Aug2 performs better than Aug1 and Aug3 performs better than Aug2 overall. The decreases in the combined WER's can not be attributed to a single speaker group, as it seems that every group decreased uniformly with only a few exceptions.

Aug1 performs better at every aspect compared to the baseline, except for adults above 65 which has a very small increase in WER. All the other groups have a decrease of a couple percentage compared to the baseline. This is the worst performing of the augmented ASR's.

Aug2 performs better in every group compared to the baseline and Aug1. The difference between Aug1 and Aug is again a couple of percent. There is again not a specific group that gains a lot more performance than the others. What is interesting in these results is that in this group the older adults also perform better compared to the baseline. Another surprising WER is that of females, that even if it already had one of the lower WER is still decreasing the most.

Aug3 combined WER improves a bit compared to Aug2, it has the lowest combined score out of all systems. Most of the scores are lower than Aug2 and therefore also lower than Aug1 and the baseline. Only children ages 12-16 is a bit higher than Aug2. However one big outlier is that the conversational speech actually has a higher score than the baseline. When considering that there is more read speech than conversational, one can imagine that this might influence the combined score.

Results Combined Augmented ASR

The final ASR AugC which has all the data of the three Aug's combined, did not give such good results. Even though it has double the training data compared to these systems, see: table 3. Most of the scores are better than the baseline with a couple percent. The conversational speech degraded even more compared to Aug3. It was expected that increasing the data with this much would give the best results. However frequency perturbation might have a diminishing return after so much data.

All in all the results did manage to answer our research questions. Our sub-questions: Can we get a WER (word error rate) better than the original baseline for the three speaker groups of children, adults and older adults? All have been answered, they all have been improved compared to our baseline. Children aged 7-11 and Adults have been improved most confidently. The other two children groups have also been improved but a little less. Older adults has only been reduced with less than one percentage if we consider Aug2 instead of Aug3 to remove the uncertainty of the increased WER of the conversational speech in Aug3. With these questions answered and looking at the combined results we can say with certainty that we can improve ASR performance on JASMIN Flemish Dutch data using data augmentation.

4 Discussion and Conclusions

To conclude frequency perturbation manages to improve the ASR performance on Jasmin Flemish Dutch data slightly. The WER of the augmented ASR systems all had a better performance compared to the baseline. The baseline was on par with the results from previous research, with the big exception of conversational speech being worse in this baseline. Almost every speaker group had a lower WER compared to the baseline, with the major exception of the ASR Aug3 using the (-36, 36) dB gain, which has a worse WER for conversational speech. Other than that it seemed that the more dB were cut or boosted the more each speaker group increased performance. Even though the performance is still lacking compared to the other Dutch dialects [1], it was an improvement none the less.

Improvements to this research could be made with training the baseline, to include only native people in the training set in order to improve the WER. Since the level of the non-natives is relatively low there is a good chance that they are degrading the ASR system. This could make it easier to improve upon this baseline since the scores are lower. Another improvement would be to try more decibel ranges, for example higher and lower ones or ones that only boost or ones that only cut. The expectation is that the improvement rate would still be around a couple of percentage increase but the probability that between the 3 augmented ASR systems it contains the best one, is of course small. One last improvement that could be made is to try out more data augmentation techniques or a combination of different techniques. There is a big chance that there are data augmentation techniques that will perform better than frequency perturbation.

5 Responsible Research

Lastly, it is important to handle the ethical aspects that affect this research. The research should be reproducible and no bias should be created.

The experiment in this research can be reproduced by the steps given in the experimental setup section. The Jasmin corpus is publicly available and the data split is given so this can be easily recreated. The data augmentation technique is described in detail so this can also be recreated. One aspect to keep in mind when reproducing this experiment is that the data augmentation technique randomizes the amount of decibel boosted or cut in between the range it gets. This means that results between different runs will slightly vary.

The data has been split in such a way that there are no forged results. All the different speaker groups are split evenly and there no speakers that overlap. So there will be no bias towards a specific group or a speaker. This makes it so the research has been conducted in a responsible manner.

References

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

[2] M. Najafian, *Acoustic model selection for recognition of regional accented speech*. PhD thesis, University of Birmingham, 2016.

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

[4] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, "Data augmentation for low resource languages," in *INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association*, pp. 810–814, International Speech Communication Association (ISCA), 2014.

[5] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[6] C. Cucchiarini, H. V. Hamme, O. v. Herwijnen, and F. Smits, "Jasmin-cgn: Extension of the spoken dutch corpus with speech of elderly people, children and nonnatives in the human-machine interaction modality," 2006.

[7] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, p. 21, 2013.

[8] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[9] N. Oostdijk *et al.*, "The spoken dutch corpus. overview and first evaluation.," in *LREC*, pp. 887–894, Athens, Greece, 2000.

[10] 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," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, Dec. 2011. IEEE Catalog No.: CFP11SRW-USB.

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

[12] M. Gales, S. Young, *et al.*, "The application of hidden markov models in speech recognition," *Foundations and Trends® in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.