Evaluating the Effect of SpecSwap for Purposes of Improving WER Performance of the Western Dutch Region Using the JASMIN-CGN Dataset

Alves Marinov
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

A problem prevalent in many modern-day Automatic Speech Recognition (ASR) systems is the presence of bias and its reduction. Bias can be observed when an ASR system performs worse on a subset of its speakers compared to the rest rather than having the same overall generalization for everyone. This can be seen by using Word Error Rates (WER) as a metric. Depending on the ASR system in question the type of bias differs. However, techniques have been proposed and shown to succeed in reducing WER, and subsequently bias, by the use of data augmentation techniques for the recorded speech. These techniques perturb the audio in a certain way. Afterward, it is added to a model's training set and the model is retrained with the added data. One such technique is SpecSwap. This paper explores how using SpecSwap affects the WER performance of a hybrid-model ASR system using the JASMIN-CGN dataset's West-Dutch region. For comparison, a state-of-the-art data augmentation technique, VTLP, was also used, which has been shown to be effective in other cases. The experiments both led to a consistent WER increase. Therefore it was concluded that the data provided for the region was too little for the augmentation policy to be effective in any of the subcategories or in the overall performance of the system. However, SpecSwap shows potential in mitigating the widely discussed gender bias in ASR systems by reducing the difference between male and female speakers' WER.

Index Terms: speech recognition, bias, bias reduction, speech augmentation, WER, SpecSwap

1 Introduction and Research Topic

For the development of Automatic Speech Recognition (ASR) systems for real-life applications researchers are heavily reliant on the existence and correctness of Large Speech Corpora (LSC), big data sets of recorded speech used to train the speech recognition models. These sorts of applications are very data-driven and data-consuming. As the demand for Human Language Technology (HLT) increases so does the variety of people exposed to and using those systems. From there it follows that those HTLs will have to be able to handle a multitude of accents, dialects, and pronunciations. This in turn drives the need for more data with wider coverage and better accuracy across everyone. This drive for an overarching general model is also paired with the prevalent problem of bias in these systems.

Bias in an ASR system exists when a system consistently performs worse on certain speaker groups. One of the metrics used for that quantification is the Word Error Rate (WER). Depending on the systems used, the corpora applied, and the methods of training, different types of biases can be observed. The one under most scrutiny and discussion recently is the gender bias, where in some cases [1] male speakers' voices are better recognized by the system while in others [2] female voices have a lower WER. However, these are not the only types of biases that are prevalent in these systems. Studies [3] [4] have shown that both children and the elderly have drastically worsening WER when it comes to certain age ranges. [1] comments that one possible reason for bias appearing could be the training set used for the models. But while that might account for the presented gender bias, [4] show that even with a balanced data set results for some speakers are consistently worse.

For the Dutch language there exists the Spoken Dutch Corpus (or Corpus Gesproken Nederlands;CGN) [5] giving researchers the ability to further develop and improve upon ASR systems. However, it is comprised of adult native speakers and does not take into account children, the elderly, or non-native speakers. [6] also reflect on the lack of representation in the corpus when it comes to the different Dutch regions and accents as this could limit its application to the wider population. [6] present the JASMIN-CGN and state that it is meant as an extension to the already available CGN. The JASMIN Corpus also provides information for the various regions the speech data was collected from, participants' age ranges, genders, and other important distinguishing factors. This gives researchers the opportunity to work with a different segmentation of the data and provides them with other crucial parameters that could influence recognition scores and further help assess and reduce biases. This paper will focus on the West-Dutch region provided by the JASMIN corpus and the dialects present within it.

Research into quantifying bias and degradation of ASR performance for the Dutch language using the JASMIN corpus and the original CGN was done by Feng et al. [7]. They outline the different influences and factors contributing to this lack of universal recognition using the provided corpora. By training on the original CGN and testing against the JASMIN corpus they outline the differences in WER of many different subgroups. This paper takes a similar approach to speaker segmentation as outlined in Section 2.

Two ways in which bias can be mitigated in an ASR system are the following. One is to create a distinct model for each area where the original one is not performing well. However, this drives development away from being able to use one generalized model. The second approach is to add more data. It also has two main ways of achieving it. The first is to provide more speech recordings and further labeling, which is a time-consuming and resource-intensive process. The second option is to use data augmentation techniques. They produce more data using the data set already available, which when added to the original training set will improve the performance of the ASR system. This second method is even more favorable in low resource environments where large amounts of new data are not readily available or easy to produce. This is the case with the JASMIN corpus, which provides about 95 hours of speech recordings. Those are with the inclusion of pauses and gaps. Additionally, the West-Dutch region has the smallest portion of available speech data.

Several data augmentation techniques [8] [9] [10] [11] have shown that they can be used for the purposes of improving the overall WER performance of a system, even in low resource environments. For the purposes of this research, several state-of-the-art data augmentation techniques [12] were considered. Among them are Speed Perturbation, SpecAugment [9], and SpecSwap [8]. Additionally, within the research team the approaches of frequency perturbation, Vocal Tract Length Perturbation (VTLP) [13] and pitch shifting were considered. This paper will mainly discuss the effects of the SpecSwap [8] policy when applied to the speaker recordings of the West-Dutch region in the JASMIN corpus. Therefore the focus will be on answering the following research questions:

- Does SpecSwap improve the Word Error Rate (WER) performance overall for the speakers from the West-Dutch region of the JASMIN corpus?
- Does SpecSwap improve the Word Error Rate (WER) performance for teenager/elderly speakers from the West-Dutch region of the JASMIN corpus?

The rest of this paper is structured as follows. Section 2 delves into the methodological approach used in answering the questions as well as the data augmentation technique applied. In Section 3 the experimental setup and results will be discussed. Section 4 will reflect on the future work that can be done as well as draw conclusions based on the results. Section 5 touches upon the topic of conducting ethical research. Section 6 includes the author's acknowledgments. Following that are the references used in writing this paper.

2 Methodology and Data Augmentation

This section first looks at the structure of the JASMIN corpus and the West Dutch region. Following that, the SpecSwap augmentation technique is discussed.

2.1 JASMIN-CGN and the Western Region

The JASMIN-CGN [6] expands upon the originally created CGN [5] by introducing about 95 hours of spoken Dutch recordings collected from Dutch natives as well as non-natives. Those in the native category include children, teenagers, and the elderly, whereas the non-natives a comprised of two groups - children and teenagers. With the JASMIN corpus, we can try to address the problem of lack of representation of certain dialects, age groups, and regions by using the speech data provided. The regions that are present in the corpus are West-Dutch, Transitional, Northern, Southern, and Flemish. Further separation of the data can be found in the two types of recordings the JASMIN corpus provides - read speech, where the speaker is reading from a script, and conversational speech, aimed at simulating dialogue by means of Human-Machine Interaction (HMI).

This paper mainly concerns itself with speakers from the Western region. According to the corpus documentation the West-Dutch region is comprised of South-Holland, excluding Goeree Overflakee, North-Holland, excluding West Friesland, and West Utrecht, including the city of Utrecht. However within the corpus only speakers from North-Holland are present. Furthermore, using the JASMIN-CGN [6] documentation, the labels associated with the specific region, and subsequently the speaker IDs that fall under that category, were found. After finding the appropriate labeling for the data that was needed the next step was to segment only those speakers from the complete data set by using their IDs. Each speaker's full spoken data was aggregated and summed up giving the total amount of recorded speech time. This is the time the speakers have spoken excluding longer pauses and gaps in the recordings. That summed up to 271.7 minutes or about 4 hours and 31 minutes. The data available for the West-Dutch region is also comprised of 38 speakers, 20 of which are teenagers and the remaining 18 are elderly. 16 of the speakers are male and among them 11 are teenagers and the other 5 are elderly. The remaining 22 speakers are female of which 9 are teenagers and 13 are elderly. The research conducted takes all listed categories into account and will analyze both read and conversational speech, male and female speakers, and children and elderly speakers.

2.2 Data Augmentation Technique

The data augmentation technique chosen and implemented by the author is SpecSwap, whereas the rest of the techniques mentioned previously are either implemented or directly applied by his research colleagues and further elaborated upon in their own research papers [14] [15] [16] [17].

The SpecSwap[8] method uses two ”basic” [8, p. 581] ”computationally cheap” [8, p. 581] augmentations. Those being frequency swapping and time swapping which occur directly on the spectrogram level. These operations can be considered very similar to image processing, before the spectrogram is converted back into audio form. [8] tell us that given parameters F and T, blocks on the corresponding axis, frequency and time respectively, are swapped with each other. The blocks that are to be swapped are chosen from two non-overlapping segments of the spectrogram based on the input parameters. Within those segments, the blocks are picked from a uniform distribution. A visual representation of how the technique works can be seen in the comparison of Figure 1 and Figure 2. In the spectrograms, the louder segments are bright yellow and orange while the quieter ones are marked with darker colors or fading to black. Figure 1 is an unmodified speech file from the JASMIN corpus [6] while Figure 2 showcases the results after applying both frequency swapping and time swapping. We can observe that the overall structure remains the same, however, the quiet spot in the middle is swapped out for a segment at the beginning and two blocks of the frequency band have also been swapped with each other. The augmentation technique does not influence the duration of the provided speech recordings. After performing SpecSwap the augmented data is added to the training set of the baseline. This gives us double the amount of training data and is the main audio augmentation technique this paper discusses.

It is important to note that Song et al. [8] introduce and apply SpecSwap in an end-to-end [18] system. However, testing the method on one such system is outside the scope of this research. Its performance on a hybrid model ASR will be further discussed and evaluated in the next section.

3 Experimental Setup and Results

This chapter follows the steps for creating the training and testing data. Following that, the toolkit used for the language model is discussed as well as the baseline WER. The augmentation technique and parameters are elaborated upon and the results are presented in the end.

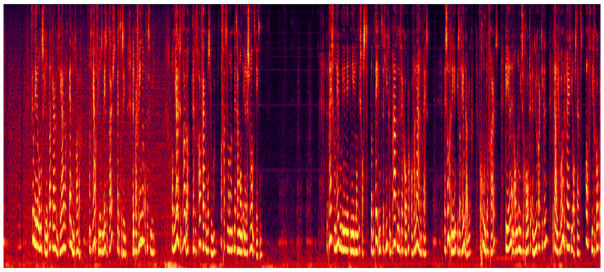

Figure 1: Mel spectrogram of one of the original speech files

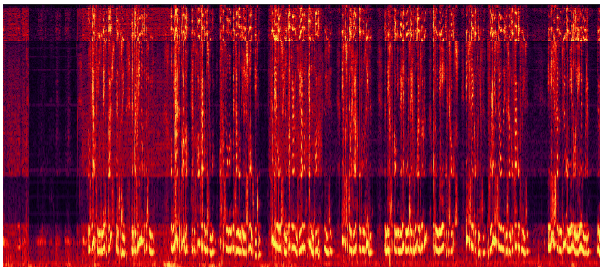

Figure 2: Mel spectrogram of the same file after SpecSwap was applied both in the frequency and time domain

3.1 Data Preparation

After the speakers of the West regions are known and their total speaking time has been computed they are separated into two groups for training and testing. The training set is used to train the model while the test set's purpose is to evaluate it by means of WER percentages. The lower the WER the better the ASR system is performing. WER is computed by summing up three categories of errors and dividing that number by the total amount of words spoken. The three categories in question are insertion, deletion, and substitution. Insertions are words that were detected but are not present in the original text or recording. Deletions are words that were not detected and therefore are missing or are deleted from the final recognition. Substitutions occur when words are recognized but there is a phoneme difference between the original and the recognized one.

To achieve a proper unbiased separation of the data equal proportions of each speaker group should be present in the training and testing sets. Additionally, we cannot have a speaker be in both sets at once. Therefore we separate them first by speaker ID and then by the total recorded speech time. The training set is approximately 80% of the regional speakers and the test one 20%. This amounts to approximately 217 minutes of training data and 55 minutes of testing data. The full separation between train and test set can be observed in Table 1 and Table 2. The speakers from the West-Dutch region in the test set have the following labels as present in the JASMIN-CGN documentation: N000{166, 175, 179, 184} and N1000{47, 77, 79}

Table 1: Distribution of speakers' time in train set in minutes

	Teenagers	Elderly	Total
Male	53.5	35.4	88.9
Female	42.5	85.2	127.7
Total	96.0	120.6	216.6

Table 2: Distribution of speakers' time in test set in minutes

	Teenagers	Elderly	Total
Male	13.5	9.6	23.1
Female	10.4	21.6	32.0
Total	23.9	31.2	55.1

3.2 Speech Recognition Setup and Baseline WER

All model training, testing, and speaker extraction was done on the supercomputer provided by the Delft High Performance Computing Centre [19]. Additionally throughout the course of this research, the Kaldi [20] toolkit was used for training and testing the data. It is an open-source speech recognition toolkit. For the training and testing of the ASR system a GMM-HMM hybrid acoustic model was used in conjunction with a 3-gram language model and a lexicon. However, due to time and resource limitations this was the only model that was used for the purposes of ASR. After training the model with the previously described parameters and data separation a baseline WER of the West-Dutch region was achieved. That was used for future comparisons after augmentation was performed. The initial WER achieved for the Western region was 25,84%. Further breakdown of each subcategory of speakers can be seen in Table 3.

Given that the data available for the West region was the least compared to the remaining ones the language model could not be trained solely on it. Therefore the text provided during training was the one available for the complete corpus without the text that corresponds to the test data, as to avoid any skewing of the results.

3.3 Audio Augmentations

For the SpecSwap [8] policy augmentations were performed with parameters as close as possible to the ones stated in the relevant paper. The optimal parameters stated in [8] were F=7 and T=40. Therefore swapping blocks was done both in the frequency and time domains. Additionally, for the frequency domain of the spectrogram, the researchers used 40 mel-banks which was also the case for the set of audio augmentation applied. For the time domain, the measure of seconds was used.

There was no official implementation of SpecSwap available so for the current experimental setup it had to be implemented from scratch. This was done in Python 3.8 with the help of several packages, the main one being ”librosa” [21]. Librosa was used to convert the speech recordings into spectrograms and back again into audio after the augmentations were performed. Others include ”sounfile” for reading and writing the audio data, ”os” for acquiring all the files, and ”numpy” for simulating the uniform distributions needed to choose the blocks for swapping.

Additionally, audio augmentations were performed using the VTLP [10] method as implemented by Zhlebinkov [14]. This was done to compare the results of SpecSwap with an already established method that provided favorable results [14] when used by others in the research team for their own regions.

3.4 Results

The results obtained from the initial baseline and the audio augmentations are in terms of WER percentages. We can observe the produced WERs in Table 3.

The ”Baseline” column contains the original JASMIN corpus data for the West region. The ”SpecSwap” column shows the results of applying the augmentation technique to a copy of the training data and adding that to the previously created training set. Similarly, the ”VTLP” column represents the results of applying that augmentation and adding that data to the training set of the baseline.

Table 3: WER percentages for the baseline model, results after applying SpecSwap and retraining it, and results after the application of VTLP and retraining the model, broken down by category

	Baseline	SpecSwap	VTLP
Combined	25.84	37.37	28.17
Male	23.94	36.70	26.2
Female	27.32	37.02	29.60
Teenagers	9.33	18.78	9.95
Elderly	41.21	54.3	45.21
Read	15.53	25.73	16.86
Conv	61.17	72.03	64.27

Observing the results of the baseline we notice that the difference between the male and female speakers' WER is about 3% with the female voices performing slightly worse on average than the male ones. This is despite the fact that in the distribution female speakers have significantly more time speaking compared to male speakers. This also contrasts with the findings by Feng et al. [7] where female speech was better recognised.

We can also observe the difference in the system's performance when it comes to teenage speakers, sometimes being orders of magnitude better than other subcategories. We see that the elderly speakers have significantly worse WER performance compared to teenagers despite representing the majority of total time spoken in the training set by a margin of about 24 minutes. These results are in line with the reported ones from Feng et al. [7].

The difference between read speech and conversational speech is apparent, with read speech performing about 4 times better than conversational. While their differences are not as drastic, Feng et al. [7] also point out that on average conversational speech has a higher WER compared to read speech.

By comparing the baseline with the other two augmentations it can be seen that they perform strictly worse in every category. SpecSwap has the worse performance of the two where the minimum increase in WER is about 9% for teenagers and the maximum about 13% for elderly speakers. VTLP also shows worse results, however with much smaller differences, where the smallest increase is close to 0.6% for teenage speakers and the highest is 4% for the elderly speakers. It is important to note that while VTLP mostly preserves the difference between male and female speakers it increases the difference between teenagers and the elderly as well as between read and conversational speech. On the other hand, SpecSwap reduces the difference between male and female speakers while also increasing the difference between teenagers and the elderly and mostly preserving that between read and conversational speech.

A possible reason for both of the models failing to improve the WERs is a lack of sufficient data, as the West region contains the least amount of total spoken time compared to all the other regions. To check the validity of this one final test was done. The spoken data from the Transitional region of the JASMIN corpus was added to the training data for the model. The test set remained the same. Then the model was retrained and tested. The WER results were compared to the previously obtained baseline and can be seen in Table 4. By adding more data, even one not from the region, the model performs strictly better in all categories while also preserving the general trends outlined before.

Table 4: The WER percentage for the baseline data of the West-Dutch region, as well as the same data combined with the data from the Transitional region after retraining, broken down by category

	Baseline	With Transitional
Combined	25.84	23.86
Male	23.94	22.57
Female	27.32	24.23
Teenagers	9.33	8.96
Elderly	41.21	37.85
Read	15.53	13.83
Conv	61.17	55.02

4 Conclusions and Future Work Discussion

This paper explored the effects of using SpecSwap[8] as an audio augmentation technique for reducing the WER of an hybrid model ASR system trained on the JASMIN-CGN[6] for the Western Dutch region. The results were also compared to the VTLP augmentation technique.

Looking at the initial baseline the difference between male and female speakers is rather small, about 3%. However, the differences between the age groups of teenagers and the elderly as well as the types of speech, read and conversational, were significantly higher. These baseline WER percentages show potential for bias with regard to female speakers and elderly speakers. Even though their speech data formed the majority of the training set their WER percentages were higher than their counterpart groups. In line with that, SpecSwap shows some potential for being able to reduce bias between genders as seen in Table 3.

Overall it was shown that all of the techniques used were ineffective for the provided data set. Throughout the tests, a better WER performance was not achieved for the speakers, neither teenagers nor the elderly, of the West-Dutch region of the JASMIN corpus by means of the SpecSwap augmentation technique. Both SpecSwap and VTLP gave increased WER percentages with SpecSwap performing worse overall.

Future research into the topic should focus on using the augmentation techniques on a bigger data set than the one provided. In the case that the JASMIN-CGN is not expanded for the West-Dutch region this could mean that generalization for that part of Dutch speakers is not possible and only a conclusion on a broader scale can be made. In the case where sufficient data for the Western dialects is available then the further usages and implications of different augmentation techniques can be derived.

5 Responsible Research

After conducting this research and observing the results it is also important to reflect on the ethical aspects that concern it as well as the reputability of the results. As such, this section is divided into two subsections. One deals with the ethical implication of the conducted research and the other with how to reproduce it.

5.1 Ethical Implications

Since the research conducted is closely related to how well an ASR system can interpret certain types of speech, which is strongly connected to bias, there could be implications for the broader use of HLT. However, this research deals with the concept and aim of reducing bias rather than furthering it. Additionally, there are no direct ethical implications that come from the results of this paper that can be addressed by the author.

5.2 Research Reproducibility

The core of this research is comprised of two main building blocks. One of them is the JASMIN-CGN[6] and the other is the recently proposed augmentation technique SpecSwap[8]. These are publicly available and given their respective documentation the research builds upon them by use of the steps detailed in Section 3. Following those, the complete research process can be reproduced. It is important to note that due to the different assigned probabilities in the ASR model results might have a slight variation from one run to the next. However, the general trend, proportionality, and relation between the different result groups should stay the same.

6 Acknowledgments

The author would like to thank the team's responsible professor, Odette Scharenborg, for giving us direction and guidance for the research topic.

Furthermore, the author would like to acknowledge the work of his complete research group. With special thanks to fellow team member Dragoș Bălan for his inspiring work, dedication, and persistence throughout the course of the research, as well as Nikolay Zhlebinkov, who provided the VTLP implementation and was a valuable discussion partner in these 10 weeks.

Finally, the author would like to extend his many thanks to the team's research supervisor, Tanvina Patel, who guided us through the steps of the process, helping us with practical matters and theoretical understanding, as well as providing timely feedback, and was always available to answer any questions we had during the research.

References

[1] 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: Association for Computational Linguistics, Apr. 2017, pp. 53–59. [Online]. Available: https://aclanthology.org/W17-1606

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

[3] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[4] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, 1996, pp. 349–352 vol. 1.

[5] N. Oostdijk, "The spoken dutch corpus. overview and first evaluation," *LREC*, pp. 887–894, 2000. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2000/html/summary/110.htm

[6] 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: http://www.lrec-conf.org/proceedings/lrec2006/

[7] 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

[8] 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*, 2020, pp. 581–585.

[9] 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.

[10] 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, 2013, p. 21.

[11] 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.

[12] H.-J. Chang. Data augmentation in automatic speech recognition. [Online]. Available: https://spectra.mathpix.com/article/2021.09.00002/asr-data-augmentation

[13] 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, 2013, p. 21.

[14] N. Zhlebinkov, "Improving northern regional Dutch speech recognition by adapting perturbation-based data augmentation," 06 2022.

[15] D. Bălan, "Evaluating the use of frequency masking on a hybrid automatic speech recognizer for transitional Dutch accent of JASMIN-CGN corpus," 06 2022.

[16] A. Mesic, "Evaluating the use of pitch shifting to improve automatic speech recognition performance on Southern Dutch accents," 06 2022.

[17] N. Sweijen, "Improving asr performance on JASMIN Flemish Dutch data by performing frequency perturbation," 06 2022.

[18] S. Wang and G. Li, "Overview of end-to-end speech recognition," in *Journal of Physics: Conference Series*, vol. 1187, no. 5. IOP Publishing, 2019, p. 052068.

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

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[21] "librosa: audio and music processing in python." [Online]. Available: https://librosa.org