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

TU Delft

Author: Alves Marinov (a.marinov@student.tudelft.nl) - Supervisor: Tanvina Patel (t.b.patel@tudelft.nl) - Responsible Professor: Odette Scharenborg (o.e.scharenborg@tudelft.nl)

1. Background & Problem

- Automatic Speech Recognition (ASR) systems rely on large speech corpora
- Current Dutch corpora are the Corpus Gesproken Nederlands(**CGN**) [1] and **JASMIN-CGN** [2] - an extension to the CGN by addition of more regions, dialects, age ranges, and non-native speakers, with its West-Dutch region currently having the least amount of data
- **Bias** has been shown to be present in the CGNs [3], expressed by **higher word error rates** (WER) for certain speaker groups
- **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**
- Different **audio augmentation techniques** [4] have shown promising results in **decreasing WER** and addressing the issue with systems requiring large amounts of data under certain conditions and parameters
- One of those techniques, recently introduced, is **SpecSwap** [5]
- However, it is yet unclear how this method **performs on a GMM-HMM hybrid system** (Figure 1) for the Dutch language

2. Research Questions

- Does SpecSwap improve the Word 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?

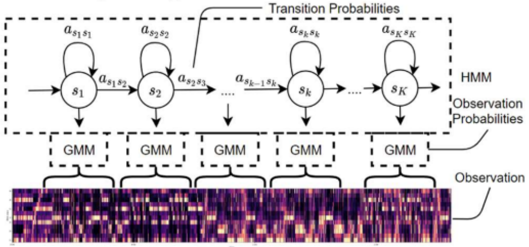

Figure 1: GMM-HMM mixture model[6]

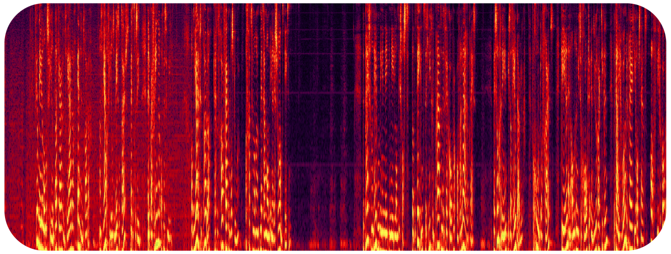

Figure 2: Unaltered speech recording spectrogram from the JASMIN corpus

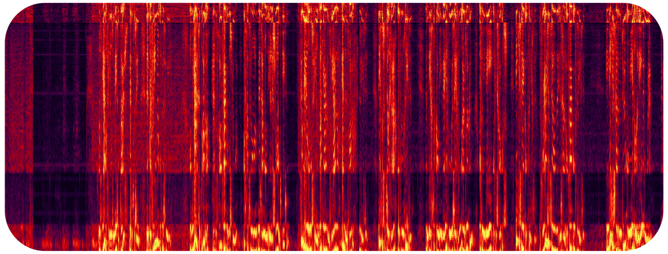

Figure 3: Spectrogram of the speech recording after applying SpecSwap

3. Method

1. Extract West-Dutch region speakers' recordings from the JASMIN corpus
2. Split the data into 80/20 training/testing sets based on speaker IDs and total time spoken, excluding gaps and pauses
3. Train and test the model to obtain baseline WER for the West-Dutch region
4. Perform data augmentation by means of SpecSwap (Figures 2 and 3 show spectrograms of an unaltered file and the same file after performing the augmentation respectively) and one by means of VTLP [7] for comparison
5. Train and test the model for both augmentation cases
6. Validate the results by training the model with the addition of the data from the Transitional region of the JASMIN corpus and testing it on the same data

4. Results and Conclusions

- Both SpecSwap and VTLP **increase the WER** for every sub-category (Table 1)
- Given that adding **more data** results in **lower WER** than the original baseline, we can conclude that the **degradation** introduced by the two augmentation techniques **is due to the very low amount of data**
- **SpecSwap**, however, **shows potential in reducing the gender bias** present in the system
- **Future work should focus on using the technique with more data**
- If more data is not made available, a specific analysis for the West-Dutch region might not be achievable

	Baseline	SpecSwap	VTLP	With Transitional
Combined	25.84	37.37	28.17	23.86
Male	23.94	36.70	26.2	22.57
Female	27.32	37.02	29.60	24.23
Teenagers	9.33	18.78	9.95	8.96
Elderly	41.21	54.3	45.21	37.85
Read	15.53	25.73	16.86	13.83
Conv	61.17	72.03	64.27	55.02

Table 1: WER percentages of the different models used, broken down by category

5. References

[1] A Reference Corpus of Written Dutch - https://lands.let.ru.nl/projects/d-coi/publs/D-COI-06-01.pdf
[2] Jasmin-CGN - http://www.lrec-conf.org/proceedings/lrec2006/pdf/254_pdf.pdf
[3] Quantifying Bias in Automatic Speech Recognition - https://arxiv.org/abs/2103.15122
[4] Data Augmentation in Automatic Speech Recognition - https://spectra.mathpix.com/article/2021.09.00002/asr-data-augmentation
[5] SpecSwap: A Simple Data Augmentation Method for End-to-End Speech Recognition - https://www.isca-speech.org/archive_v0/Interspeech_2020/pdfs/2275.pdf
[6] Study on a CNN-HMM Approach for Audio-Based Musical Chord Recognition - https://iopscience.iop.org/article/10.1088/1742-6596/1802/3/032033
[7] Vocal tract length perturbation (vtlp) improves speech recognition - http://www.cs.toronto.edu/~ndjaitly/jaitly-icml13.pdf