Using data augmentation to improve accuracy for ASR for Northern region Dutch accents

1. Important terms

Automatic Speech Recognition (ASR):

- converts speech recording to text
- often measured by **Word Error Rate** (WER)[1]
- needs data to train and be more accurate

$$WER = \frac{S+D+I}{N}$$

S - № substitutions
D - № deletions
I - № insertions
N - total words spoken

Data augmentation: Existing data can be changed using signal processing techniques (called **perturbations**) and then added as additional training data

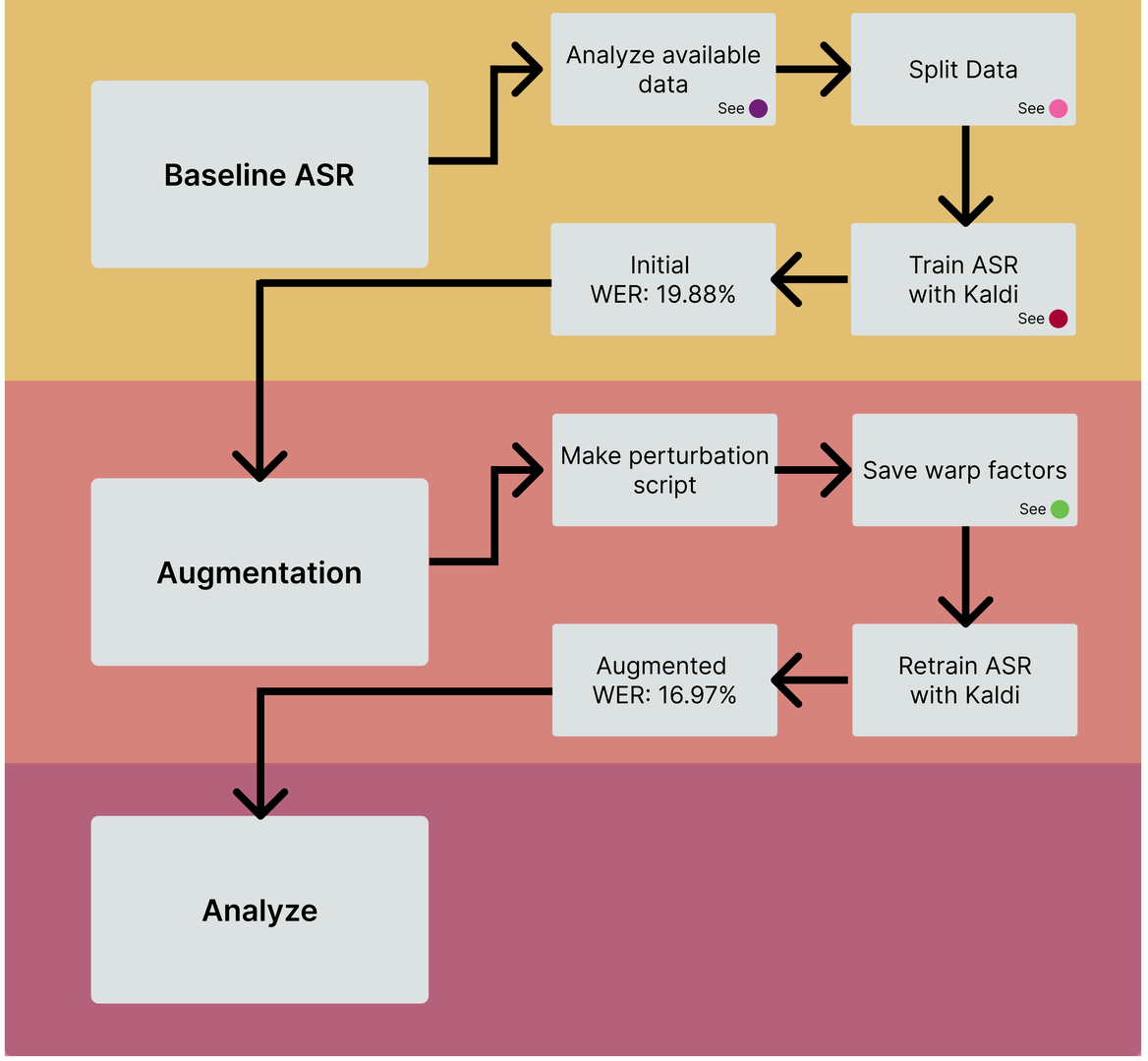
2. Goal

Explore VTLP as augmentation method, aiming to decrease WER.

VTLP (Vocal Tract Length Perturbation)[2]: entails randomly warping the frequency of each speach recording, simulating a different vocal tract

Vocal tract length on average:[3]
Children < Females < Males

3. Methods



4. Results

Comparison:

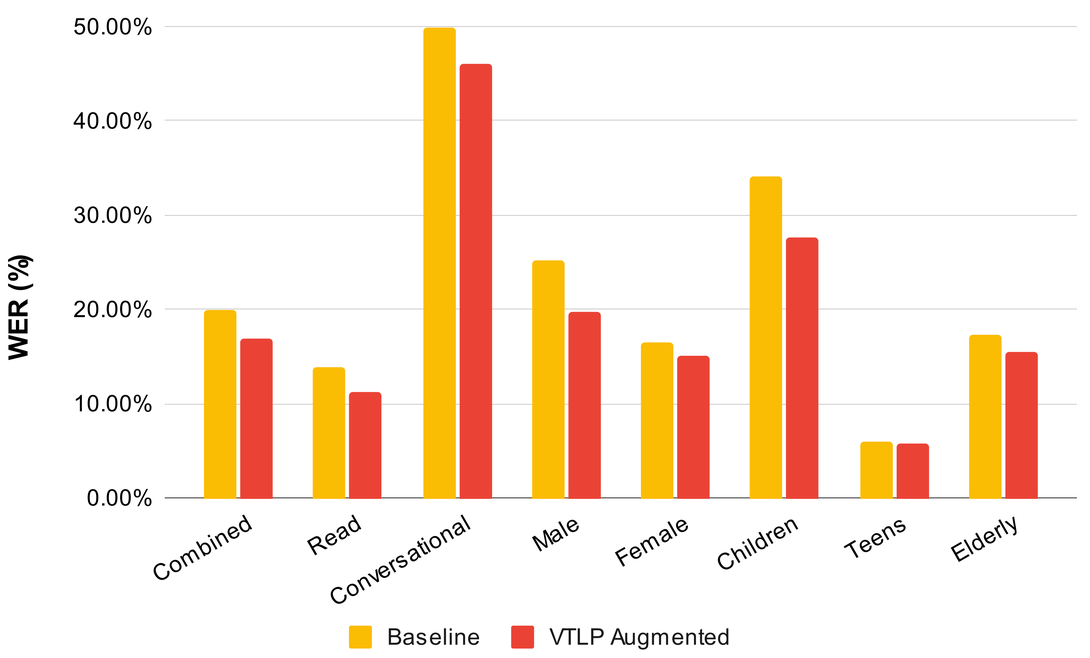
- Decrease in WER for all
- Biggest decrease in males and children

Limitations:

- Not representative of other accents or non-accented Dutch
- Unknown correlation between warp factors and VTLP efficiency

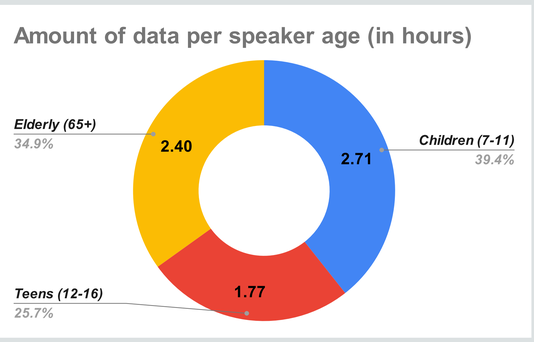
Conclusions:

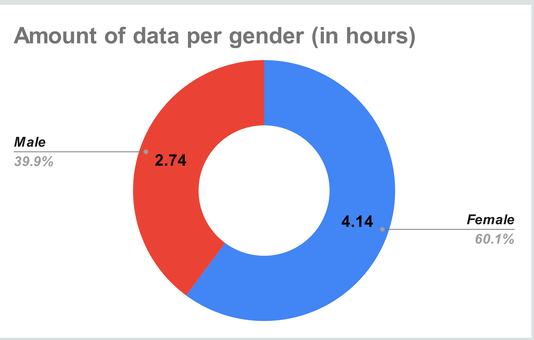
- Bias reduced
- Future: Explore VTLP based on limitations



Data analysis

JASMIN-CGN[4] - contains speech recordings of accented Dutch, annotated by gender & age





HCI (human-computer interaction) - speech recordings that simulate dialogue with machine

Read - speech recordings from a read script

Data split

- Ratio used - 80% train and 20% test
- Same speaker must not appear in both sets
- Speaker characteristics (age, gender) need to have same ratio in both sets

Train ASR - Kaldi

Kaldi - toolkit to create ASR from given `test` and `train` data sets using GMM-HMM

GMM-HMM ASR training:

- Acoustic model: Gaussian Mixture Models
- Language model: trigram - look at last 3 words
- Lexicon: what words are in the language and how they are pronounced

Reproducibility

Since warp factors are randomly generated, they need to be saved for experiment to be reproduced

Info and code provided in GitHub repo[5]

References

[1] Word Error Rate - https://en.wikipedia.org/wiki/Word_error_rate

[2] Vocal Tract Length Perturbation (VTLP) improves speech recognition - http://www.cs.toronto.edu/~ndjaitly/jaitly-icml13.pdf

[3] Morphology and development of the human vocal tract: A study using magnetic resonance imaging - https://doi.org/10.1121/1.427148

[4] JASMIN-CGN - https://aclanthology.org/L06-1141/

[5] GitHub repo with code for reproducibility - https://github.com/NZhlebinkov/research-project-2022

CSE3000 - Research Project
Nikolay Zhlebinkov
N.A.Zhlebinkov@student.tudelft.nl

Supervisor: Tanvina Patel
Responsible Professor: Odette Scharenborg

TU Delft