Using SpecAugment to develop an ASR for Transitional Dutch accent of JASMIN-CGN corpus

by Dragoș Alexandru Bălan (D.A.Balan-1@student.tudelft.nl); Supervisor: Tanvina Patel; Responsible professor: Odette Scharenborg

TU Delft

1. Background

- Automatic Speech Recognition (ASR) systems need large amounts of data → **data augmentation techniques**
- Examples of augmentation: frequency perturbation, pitch shifting, VTLP, SpecSwap, **SpecAugment (frequency masking from SpecAugment was used)**
- ASRs can be biased for specific genders[1], age groups[2] or even regional accents → **train ASR on regionally-accented data**
- Dataset used: JASMIN Dutch corpus[3]
- Available regional accents for Netherlands in the corpus: West, North, South, **Transitional**
- **Goal:** reduce bias and improve WER (Word Error Rate)
- WER = #errors(insertions, deletions, substitutions) / #words actually spoken
- Hybrid architecture used: GMM-HMM acoustic model + tri-gram language model + lexicon

2. Research question

Can data augmentation using SpecAugment improve the performance of an ASR system on the JASMIN-CGN corpus for the Transitional Dutch accent?

- Can the WER be lowered by augmenting data using SpecAugment for the JASMIN-CGN Transitional speech?
- Are there significant differences in performance between different speaker/speech categories (age, gender, conversational vs. read)?

Conversational = conversation simulation between a human speaker and a machine
Read = speech read from a script

3. Process

1. Split Transitional data into 80% train/20%test with similar distribution of age/gender between both sets
2. Train baseline model on train set then test to obtain preliminary WER
3. Augment data using frequency masking from SpecAugment (mask a frequency range of the audio spectrogram)
4. Train model with augmented data and test
5. Train 2 more models for comparisons: augmented using VTLP and Transitional+West train data
6. Compare models, analyze results, and draw conclusions

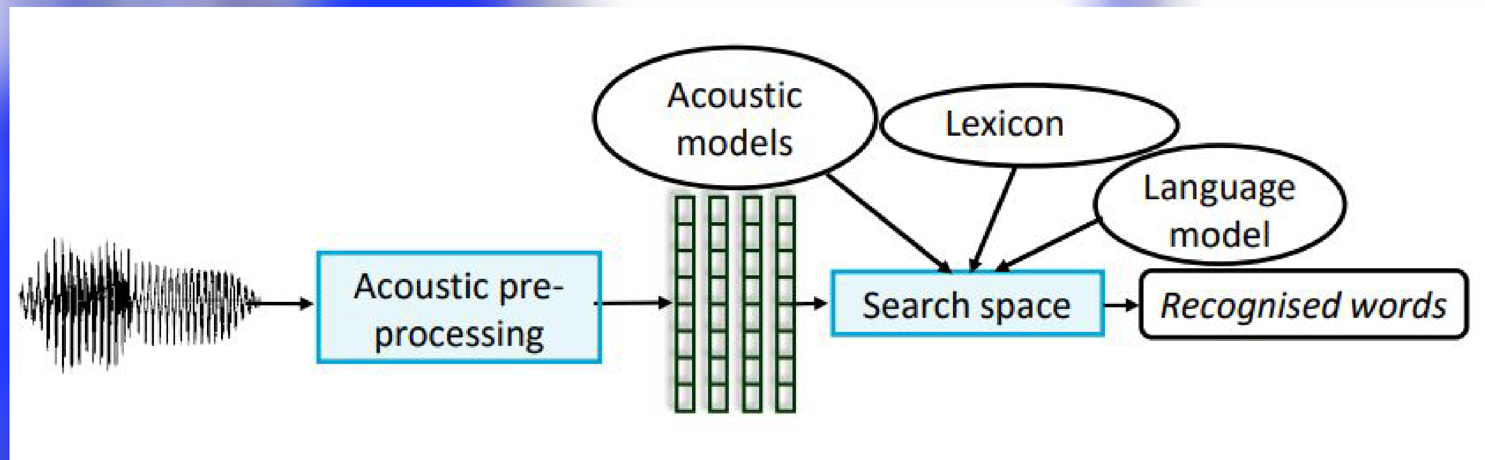

Figure 1: ASR system example [4]

4. Results

Baseline: Model trained on original Transitional data
SpecAugment: Model trained on original+SpecAugment
VTLP: Model trained on original+VTLP
Tran+West: Model trained on Transitional+West data
For all results, the smaller, the better
Overall WER: VTLP **best**, SpecAugment **worst**
Gender gap: VTLP **best**, Transitional+West **worst**
Age gap: VTLP **best**, Transitional+West **worst**
Read vs conversational: Transitional+West **best**, SpecAugment **worst**

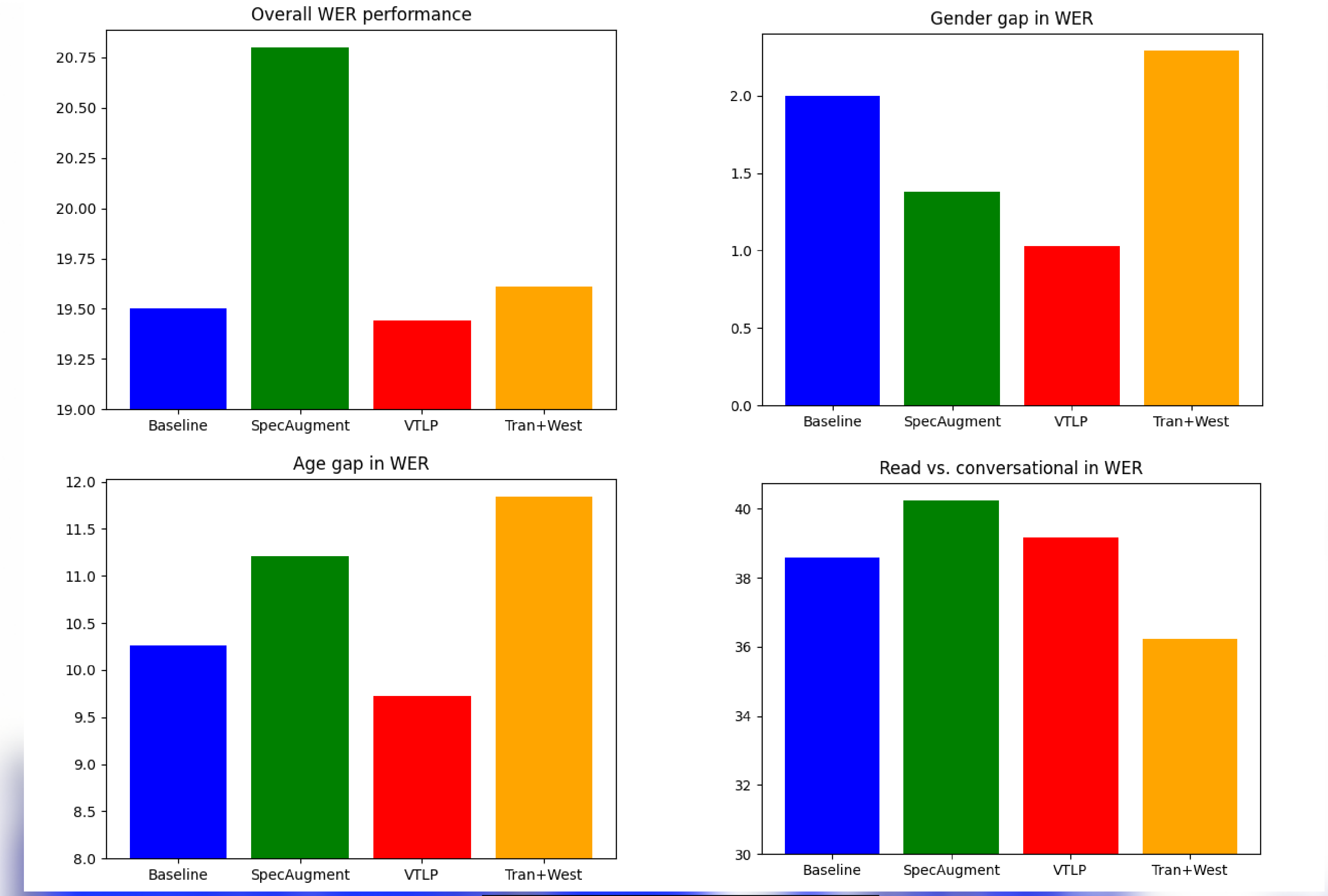

Figure 2: Results

5. Conclusions

- SpecAugment failed to reduce the WER. It also widened the gap for age and read/conversational speech
- VTLP performed the best in almost all categories
- Recommended to use VTLP instead of SpecAugment in this scenario
- SpecAugment meant for end-to-end (e2e) mainly, tested on hybrid system here → does not work well for hybrid systems and limited data

Future work:

- Test SpecAugment on entire data from JASMIN-CGN, to see if data can be an issue
- Develop an e2e system with SpecAugment+teammates' techniques, on the entire corpus

[1] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?," pp. 2205–2208, 09 2005.
[2] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying Bias in Automatic Speech Recognition," 2021.
[3] 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, May 2006.
[4] Slides by Scharenborg O., of the course "CSE2230: Multimedia Analysis"