Amar Mešić
A.Mesic@student.tudelft.nl
2022-06-22
Supervisor: Tanvina Patel
Responsible Professor: Odette Scharenborg

Evaluating the Use of Pitch Shifting to Improve Automatic Speech Recognition Performance on Southern Dutch Accents

1. Background

- ASRs are a growing technology, and rely on large speech corpora.
- Dutch language has two main corpora, CGN, and an extended JASMIN-CGN [1], which contains regionally accented data.
- JASMIN-CGN **still** has biases for different speaker groups [2]
- ASR performance is measured using Word Error Rate (WER) by measuring (substitutions + insertions + deletions) / (# words spoken).
- Researchers have been successful in using certain data augmentations to improve ASR performance using existing corpus data [3, 4].
- Many data augmentations exist, namely: SpecSwap, VTLP, PS.
- Pitch shifting (PS) is the data augmentation technique used in this project, showing potential to improve ASR performance in other studies [5].
- In this project we evaluate a hybrid GMM-HMM ASR.

2. Research Questions

Can augmenting data from the existing JASMIN-CGN corpus using pitch shifting improve ASR performance on southern Dutch accents?

- Is it possible to get an improved WER on an ASR trained with augmented Southern Dutch data from JASMIN-CGN?
- Is it possible to get an improved WER for children and the elderly on an ASR from Southern Dutch data from JASMIN-CGN?
- Is it possible to get an improved WER for non-native speakers on an ASR from Southern Dutch data from JASMIN-CGN?
- Is it possible to reduce the difference in WER for male and female speakers on an ASR from Southern Dutch data from JASMIN-CGN?

Table 1: Columns with WER from left to right; Full baseline, PS+30%, NT baseline, ±30%

WER (%)	Baseline	Pitch Shift +30%	Natively Trained (NT) Baseline	NT ±30%
Combined	43.48	44.65	60.02	45.86
Conversational	62.53	63.48	74.53	64.78
Read	37.1	38.23	54.97	39.62
Male	43.37	44.84	60.19	46.22
Female	43.29	44.31	59.83	45.42
Age Group 1	52.24	53.69	55.4	53.66
Age Group 2	19.17	21.63	14.6	25.43
Age Group 3	42.38	43	65.62	44.35
Age Group 4	41.18	41.5	72.69	41.87
Age Group 5	55.91	57.54	54.39	58.9
Native	44.47	46.38	43.47	48.02
Non-native	42.06	42.61	68.91	43.67

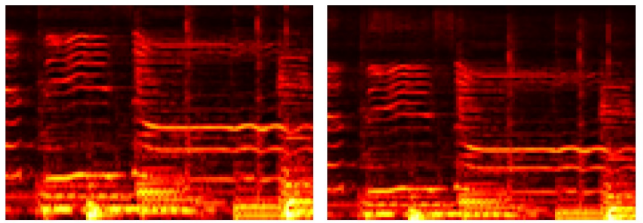

Figure 1: (left) spectrogram of original data, (right) same spectrogram with downward pitch shift of 40%.

3. Method

- Data preparation done in Kaldi: split 80% – 20% equally among gender, age, nativity.
- Southern Dutch Data: Full and Native-only baselines created using only corpus data, and WER obtained.
- Apply pitch shifting, by shifting up mel-frequency.
- Audiomentations was used to apply pitch shift to all my data.
- Augmentations applied: 30% & 50%, ±30% for natively trained ASR.

4. Results

- No improvements on fully trained baseline at all.
  - **Bias was not reduced** for any group
  - Smallest deterioration came from +30% and +50%, largest from negative pitch shifts.
  - Combining upward and downward shift did not improve performance.
  - Females performed slightly better in all test runs, but difference is not significant.
- Natively Trained ASR
  - **Improvements** in overall performance - relatively **23.6%.**
  - Of the native speakers, children improved but not teenagers and elderly did not.
  - Non-native speakers **improved** - relatively **36.6%.**

5. Conclusion

- PS on fully trained baseline - **no improvements** in WER or reduction in bias due to large presence of non-natives
- Upward pitch shift has lower deterioration
- PS on natively trained ASR
  - **Improvements** in overall performance
  - **Reduction in bias** against children and non-natives
- **No bias** between male and female speakers

Future work:

1. Other data augmentations on Southern Dutch data
2. Pitch shift on other regions of JASMIN-CGN

[1] Catia Cucchiarini, Hugo Van hamme, Olga van Herwijnen, and Felix 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, May 2006. European Language Resources Association (ELRA).
[2] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. Quantifying bias in automatic speech recognition, 2021.
[3] Shakti P. Rath Anton Ragni, Kate M. Knill and Mark J. F. Gales. Data augmentation for low resource languages. Interspeech, 2014.
[4] Matthew Baas and Herman Kamper. Voice conversion can improve asr in very low-resource settings, 2021.
[5] Mohi Reza, Warida Rashid, and Moin Mostakim. Prodorshok i: A bengali isolated speech dataset for voice-based assistive technologies: A comparative analysis of the effects of data augmentation on hmm-gmm and dnn classifiers. In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pages 396–399, 2017.