i|Home|Lab HSLU Lucerne University of Applied Sciences and Arts

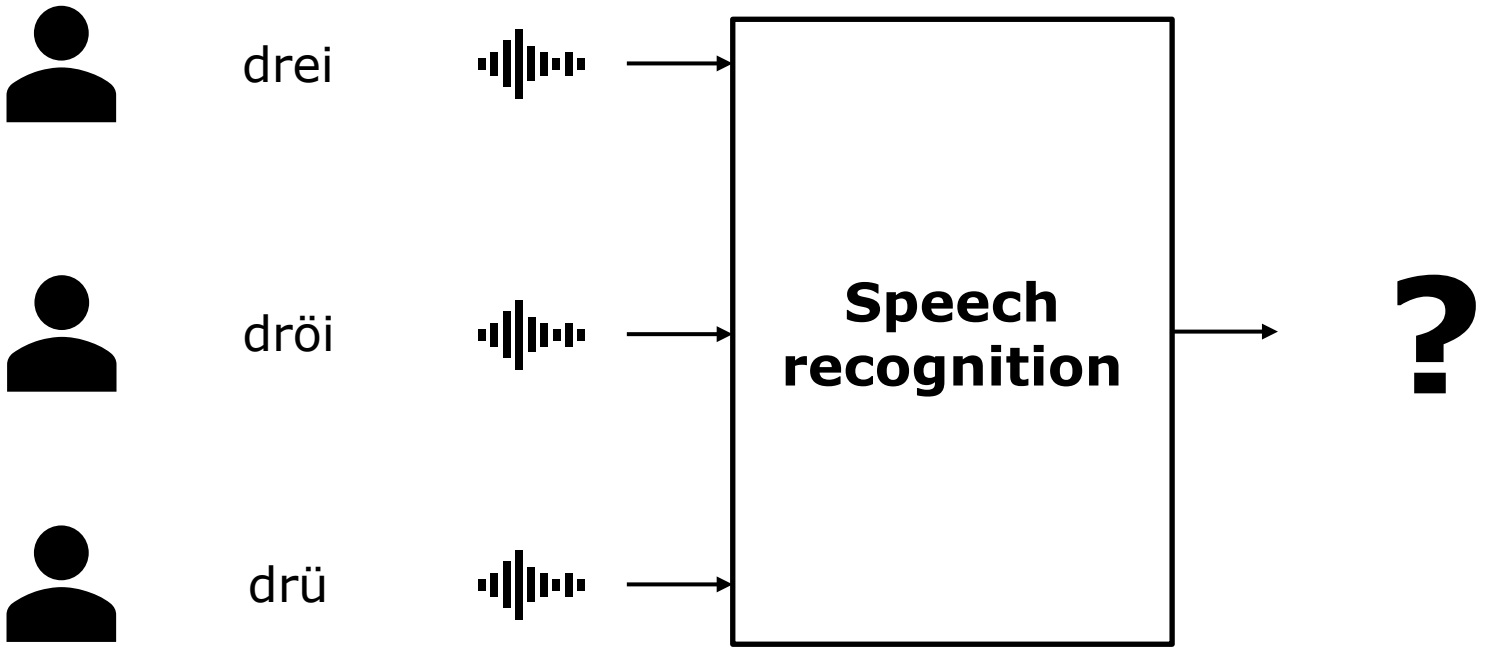
Improved Dialect Recognition by Adaptation to a Single Speaker

Manuel Vogel

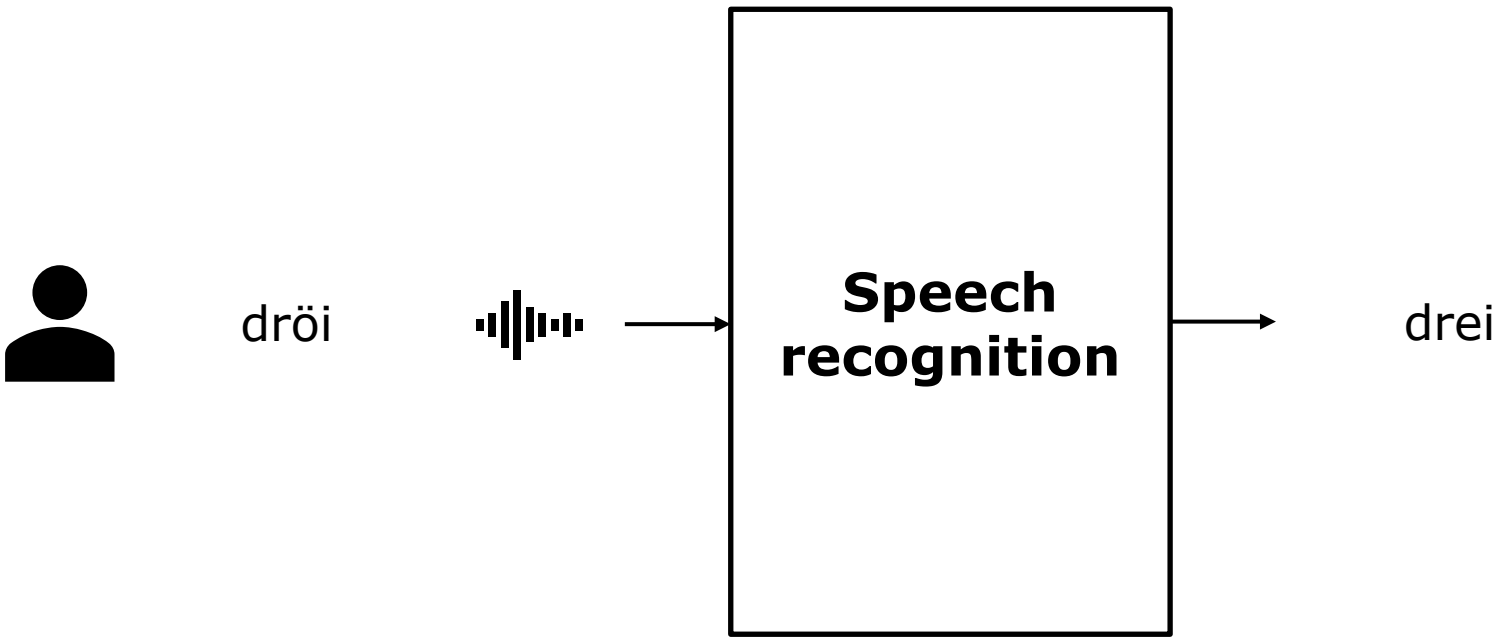
03.06.2022

FH Zentralschweiz

Motivation



Motivation

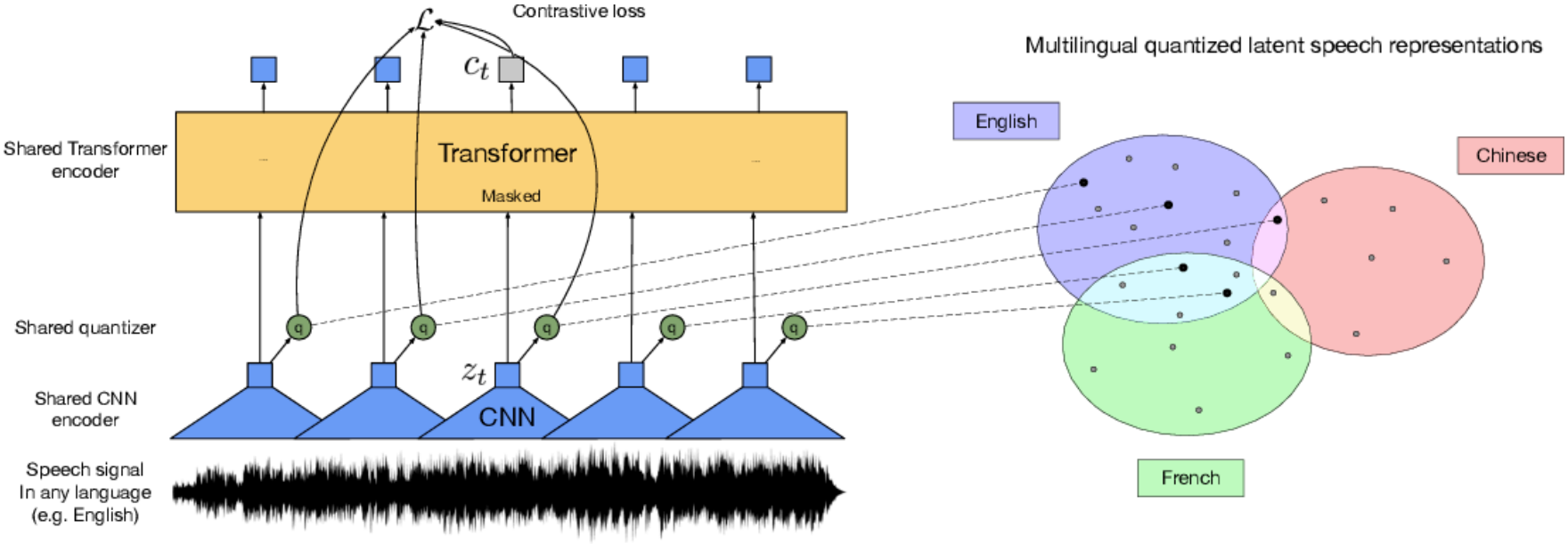


Challenges

- Continuous borders of dialects
- No standardized written form
- Desired output Standard German
- Translation step needed
- Non-applicable word boundaries, e.g “bimene” → “bei einem”
- Non-existing words, e.g “töff” → “motorrad”
- No past simple in Swiss German

Model

- wav2vec2-xlsr-53-german
- Facebook AI
- September 2020



Model – Pre-Training

- Comparable with BERT (Masked Language Model)
- Prediction of masked frames
- 53 Languages
- 56k hours audio data
- Datasets:
  - Multilingual LibriSpeech
  - Common Voice
  - Babel
- Further training on German data
- 64 GPUs used (V100)

Evaluation Metric

- Word error rate (WER)
- Most common metric in the field of speech recognition

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

- Where:
  - S = Number of substitutions
  - D = Number of deletions
  - I = Number of insertions
  - N = Number of words in the reference

Evaluation Metric - Example

		WER
Ground truth	ich danke allen mitarbeiterinnen und mitarbeitern der parlamentsdienste herzlich für die unterstützung die sie jederzeit gewähren	
Prediction	ich danke allen mitarbeiterin und mitarbeiten der parlamentsdienste herzlich für die unterstützung die sie jede zeit gewähren	0.25

$$WER = \frac{S + D + I}{N} = \frac{3 + 0 + 1}{16} = 0.25$$

- Where:
  - S = Number of substitutions
  - D = Number of deletions
  - I = Number of insertions
  - N = Number of words in the reference

Datasets

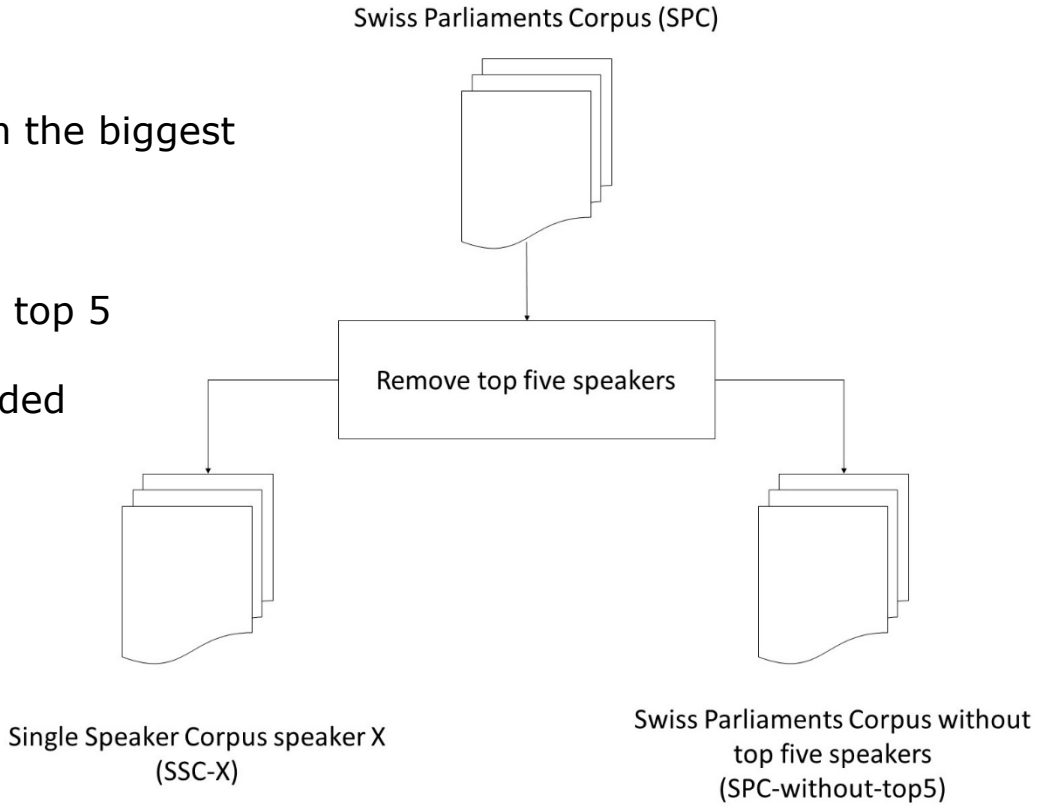
- We need:
  - Multiple Speaker
  - Single Speaker

- Based on Swiss Parliaments Corpus
  - Recordings of the parliament “Grosser Rat Kanton Bern”
  - 293h audio and 198 speakers
  - Swiss German audio
  - Standard German transcriptions

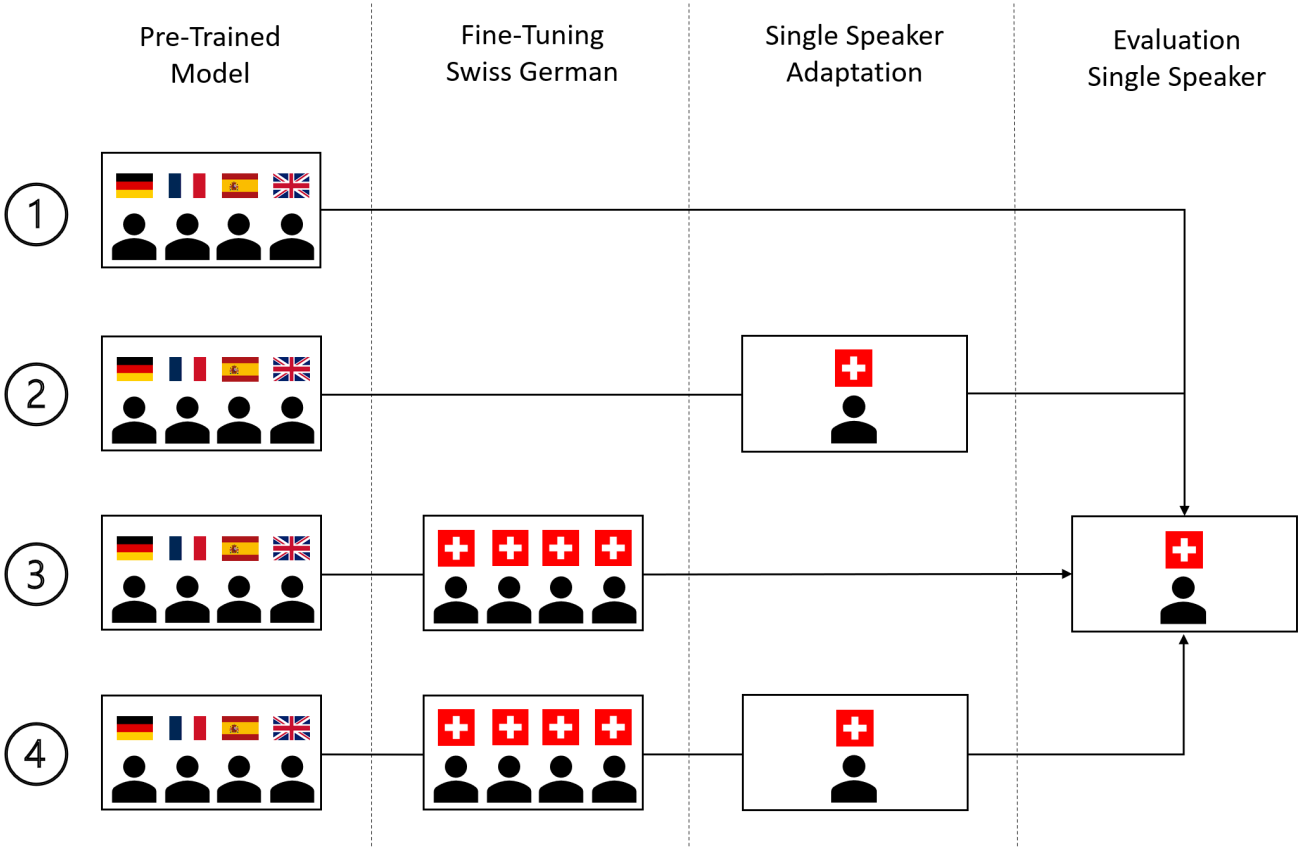
Dataset partitioning

- Multiple Speakers:
  - Remove the five speakers with the biggest amount of data in the SPC

- Single Speakers:
  - Five datasets are given by the top 5 speakers
  - One additional has been recorded
  - Limit all six to 1.5 hours



Approaches Single Speaker Adaptation



Results

Approach	Fine-Tuning	82	186	207	177	145	ext
1	-	0.90	0.89	0.82	0.89	0.86	0.82
2	SSC	0.56	0.67	0.60	0.64	0.60	0.50
3	SPC	0.31	0.42	0.44	0.35	0.38	0.44
4	SPC+SSC	**0.27**	**0.39**	**0.41**	**0.34**	**0.36**	**0.30**

SSC: Single Speaker Corpus (1.5h)
SPC: Swiss Parliaments Corpus without top five speakers

Results

Approach	Fine-Tuning	82	186	207	177	145	ext
1	-	0.90	0.89	0.82	0.89	0.86	0.82
2	SSC	0.56	0.67	0.60	0.64	0.60	0.50
3	SPC	0.31	0.42	0.44	0.35	0.38	0.44
4	SPC+SSC	**0.27**	**0.39**	**0.41**	**0.34**	**0.36**	**0.30**

Results

Approach	Fine-Tuning	82	186	207	177	145	ext
1	-	0.90	0.89	0.82	0.89	0.86	0.82
2	SSC	0.56	0.67	0.60	0.64	0.60	0.50
3	SPC	0.31	0.42	0.44	0.35	0.38	0.44
4	SPC+SSC	0.27	0.39	0.41	0.34	0.36	0.30

Further Experiments

- Increasing training dataset size SSC:
  - 82 from 1.5h to **6h** = **-2% WER**
  - 207 from 1.5h to **4h** = **-2% WER**

- Decreasing training dataset size SSC:
  - ext from 1.5h to **0.5h** = **+4% WER**
  - ext from 1.5h to **0.25h** = **+5% WER**

Analysis of Predictions – Speaker 82

	
Ground truth	in diesem sinn erkläre ich die septembersession für eröffnet

Analysis of Predictions – Speaker 82

	
Ground truth dataset	in diesem sinn erkläre ich die septembersession für eröffnet
Ground truth audio	in diesem sinn erkläre ich die session vom september als eröffnet

Analysis of Predictions – Speaker 82

	
Ground truth dataset	in diesem sinn erkläre ich die septembersession für eröffnet
Ground truth audio	in diesem sinn erkläre ich die session vom september als eröffnet
wav2vec2-german	vom
wav2vec2-german+ssc	in dem sinn erklären nie die session vom september als eröffnunt
wav2vec2-german+spc	in diesem sinn erkläre ich die session des septembers als eröffnet
wav2vec2-german+spc+ssc	in diesem sinn erkläre ich die session de september eröffnet

Analysis of Predictions – Speaker 82

		WER
Ground truth dataset	in diesem sinn erkläre ich die septembersession für eröffnet	
Ground truth audio	in diesem sinn erkläre ich die session vom september als eröffnet	
wav2vec2-german	vom	1.00
wav2vec2-german+ssc	in dem sinn erklären nie die session vom september als eröffnunt	0.89
wav2vec2-german+spc	in diesem sinn erkläre ich die session des septembers als eröffnet	0.44
wav2vec2-german+spc+ssc	in diesem sinn erkläre ich die session de september eröffnet	0.33

Analysis of Predictions – Speaker 82

		WER
Ground truth dataset	in diesem sinn erkläre ich die septembersession für eröffnet	
Ground truth audio	in diesem sinn erkläre ich die session vom september als eröffnet	
wav2vec2-german	vom	0.91
wav2vec2-german+ssc	in dem sinn erklären nie die session vom september als eröffnunt	0.36
wav2vec2-german+spc	in diesem sinn erkläre ich die session des septembers als eröffnet	0.18
wav2vec2-german+spc+ssc	in diesem sinn erkläre ich die session de september eröffnet	0.18

Conclusion and Outlook

- Improvement individual per speaker
- Dependency on Swiss German SPC dataset and its bias
- Standard German Samples
- Best results when training with SPC+SSC
- Difficulty of labeling
- Do other models show similar results?
- Behavior of other speaker identical?
- Optimal approach?
- Carefully choose evaluation metric

i|Home|Lab HSLU Lucerne University of Applied Sciences and Arts

Improved Dialect Recognition by Adaptation to a Single Speaker

Manuel Vogel

03.06.2022

FH Zentralschweiz