WP 2.2 Learners' Assessment Environments

Anssi Moisio[1]
Dejan Porjazovski[1]
Aku Rouhe[1]
Yaroslav Getman[1]
Anja Virkkunen[1]
Ragheb AlGhezi[1]
Mietta Lennes[2]
Tamas Grosz[1]
Krister Lindén[2]
Mikko Kurimo[1]

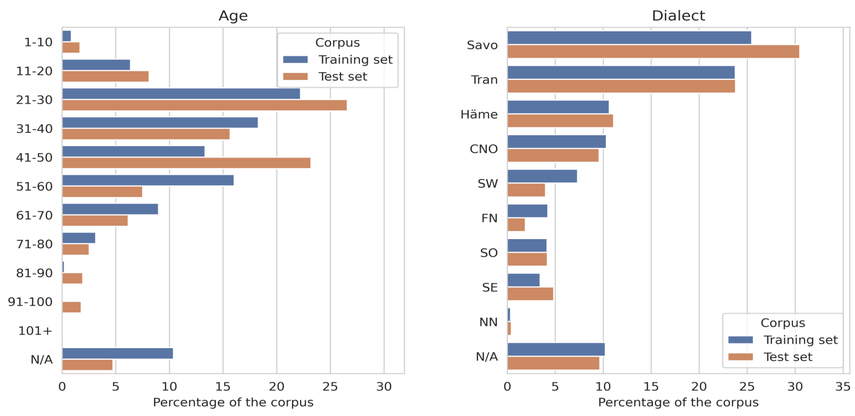
1. Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
2.University of Helsinki, Department of Digital Humanities, Helsinki, Finland

MOTIVATION

- Machine learning models trained on native speech perform poorly when evaluated on non-native speakers.
- Non-native speech characteristics:
  - Accent;
  - Limited vocabulary;
  - Grammatical mistakes;
  - Slow speaking rate;
- The non-native speech can sometimes be undetected or neglected.
- It is important to have reliable tools for identifying and properly handling the non-native speech.

LAHJOITA PUHETTA CORPUS

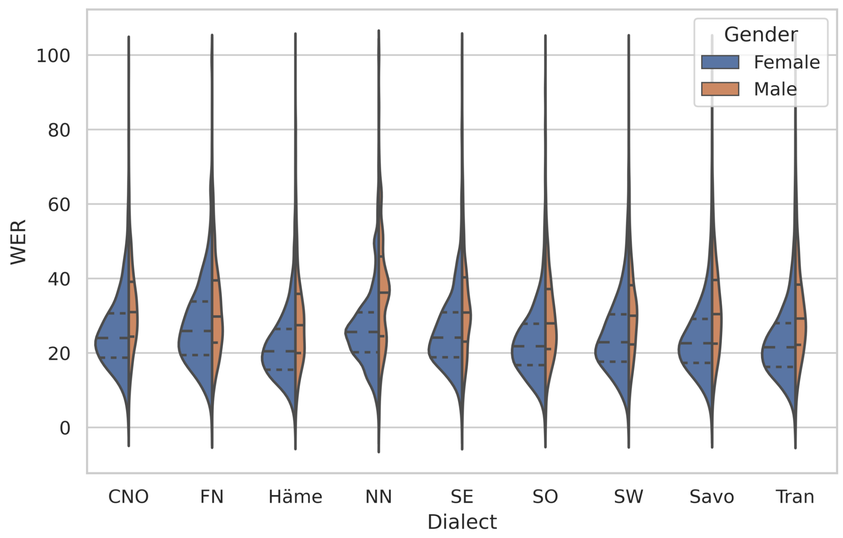
- Large-scale conversational Finnish corpus.
- Over 20000 speakers.
- Over 3200 hours.
  - 1600 hours of transcribed training data;
  - 1600 hours of untranscribed training data;
- Diverse dialects from many regions of Finland.
- Small portion of non-native speakers.



The amount of speech from various speaker groups

SPEECH RECOGNITION

- Speech recognition accuracy is lower for non-native speakers.
  - Less non-native speech to train models, only about 2% of the dataset
  - Presumably, more variation in pronunciation and grammar
- Word error rate (WER) is 31.0% for non-native speakers and 26.1% for native speakers on the Lahjoita Puhetta data.
- Dialectal variation within native Finnish speakers is not as large: the dialect group with the highest WER among native Finnish speakers got 28.8% WER.



Speech recognition word error rate per dialect. The non-native speakers (NN) have a higher WER.

METADATA CLASSIFICATION

- 4 classifiers:
  - Gender classifier:
    - Male;
    - Female;
  - Age classifier:
    - 1-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, 91- 100, 101+;
  - Topic classifier:
    - Animal friends (A);
    - Sports moments (SP);
    - Rated R (R);
    - Nature (N);
    - My surroundings (M);
    - Media skills (MS);
    - The cursed Covid (C);
    - Summer (S);
  - Dialect classifier:
    - The Southwestern dialects (SW);
    - The transitional dialects between the Southwestern and Häme dialects (Tran);
    - The Häme (Tavastian) dialects;
    - The dialects of South Ostrobothnia (Pohjanmaa) (SO);
    - The dialects of Central and North Ostrobothnia (Pohjanmaa) (CNO);
    - The dialects of Peräpohjola (the Far North) (FN);
    - The Savo dialects;
    - The Southeastern dialects and a few transitional dialects bordering on them (SE);
- Trained and evaluated on 50 second segments.
- X-vector architecture used to process the log-Mel filter banks.
- Trained audio-only, text-only, and hybrid models.
- The audio-only models were used to compare the native and non-native speakers.

Table 7 Hyperparameters of the audio encoder.

Layer	Input size	Output size	Context	Dilation
TDNN 1	40	512	5	1
TDNN 2	512	512	3	2
TDNN 3	512	512	3	3
TDNN 4	512	512	1	1
TDNN 5	512	1500	1	1
Statistical pooling	1500	3000	/	/
Linear	3000	512	/	/

Accuracy for gender classification.

Native male (262 smp)	Native female (318 smp)	Non-native male (18 smp)	Non-native female (2 smp)
85.5%	99.6%	100%	100%

Relaxed accuracy for age classification.

Native (582 smp)	Non-native (43 smp)
86.6%	83.7%

Accuracy for topic classification.

Native (347 smp)	Non-native (27 smp)
62.8%	59.2%

Accuracy for dialect classification.

Native (565 smp)	Non-native (14 smp)
41.5%	7.1%

CONCLUSION

- The conducted experiments show that the models trained on native speakers struggle to match the performance when evaluated on non-native speakers.
- This is consistent for ASR, as well as for most of the metadata classification experiments (gender classification being an exception).
- By identifying the problem, we showed that special care is needed when developing machine learning models, in order to equally represent the diverse groups that are going to use the system.