Automaattinen järjestelmä puhutun kielitaidon arviointiin

Aalto University

UNIVERSITY OF HELSINKI

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

DigiTala-projekti:
(2019 – 2023)

Aalto: ***Mikko Kurimo,*** **Ekaterina Voskoboinik, Yaroslav Getman**, Ragheb Al-Ghezi

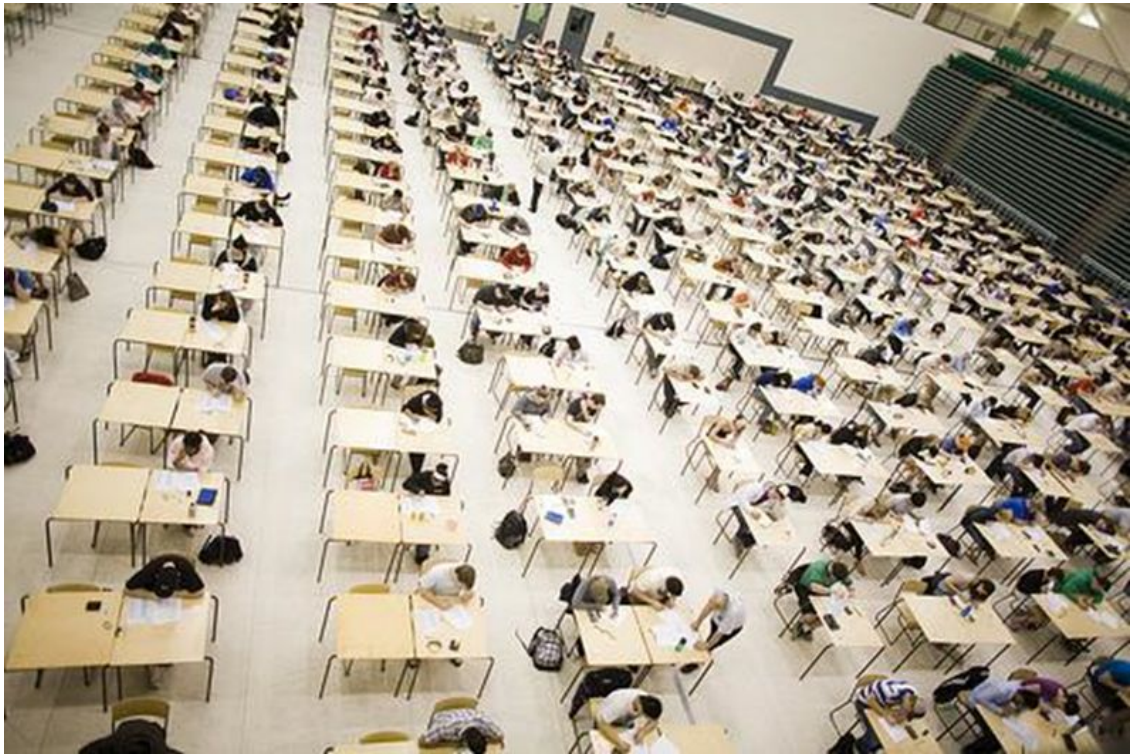
Helsinki: **Raili Hildén**, **Anna von Zansen**

Jyväskylä: **Ari Huhta, Heini Kallio, Mikko Kuronen**

SUOMEN AKATEMIA

Summary in English

- This presentation describes how we implemented an automatic speaking assessment system for second language learners
- We developed two systems, Finnish and Swedish
- We made tools for offline assessments and a Moodle-based online system that we demonstrate today



The 4+4 steps to implement this

Aalto University

Preparing the data:

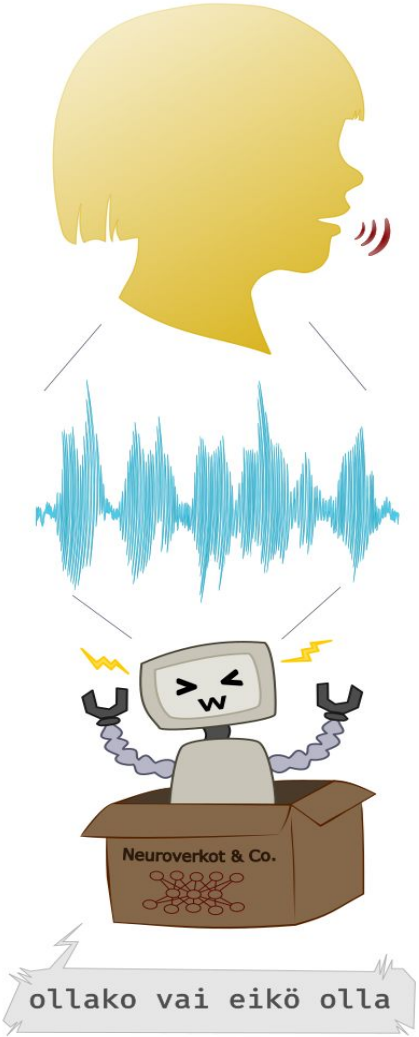
1. **Designing speaking tasks** for data collection and speaking assessment for the target groups
2. **Collecting training data** for automatic speech recognition (ASR) and Evaluators for the analytic dimensions and the holistic overall score
3. Designing the transcription guidelines and **transcribing the speech data**.
4. Designing the rating scales and organizing the **rating of the student responses** using the teachers' Moodle system that was developed for the project

Preparing the systems:

5. **Training and testing the ASR and Evaluators**.
6. **Implementing the server** back-end for online ASR transcript and evaluation scores.
7. **Implementing a demonstration** system front-end for Moodle.
8. **Testing and collecting feedback** from the test-takers and teachers.

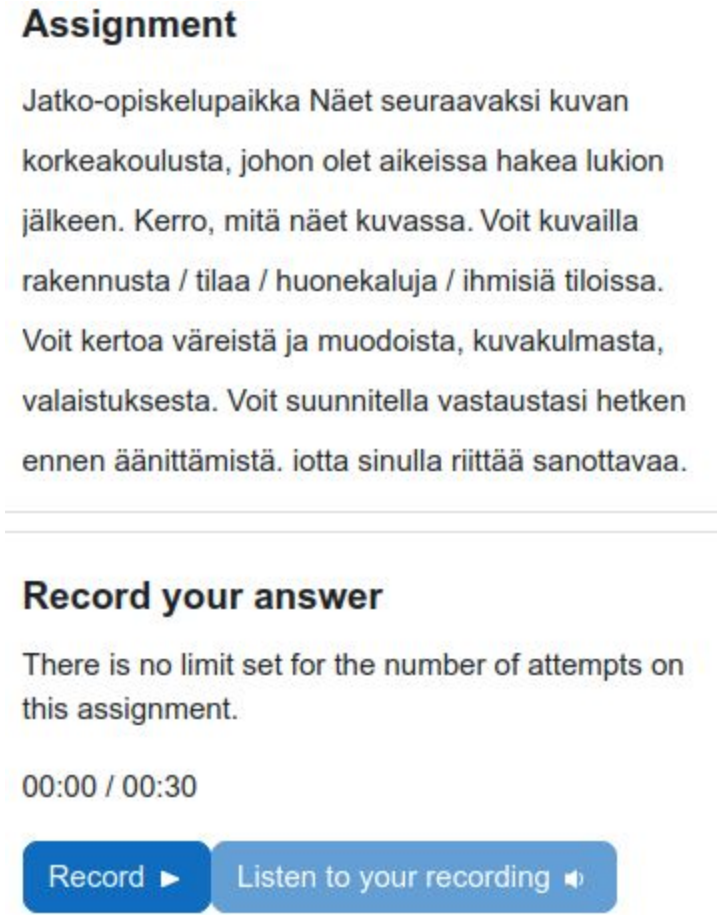
Mitä kone tuottaa?

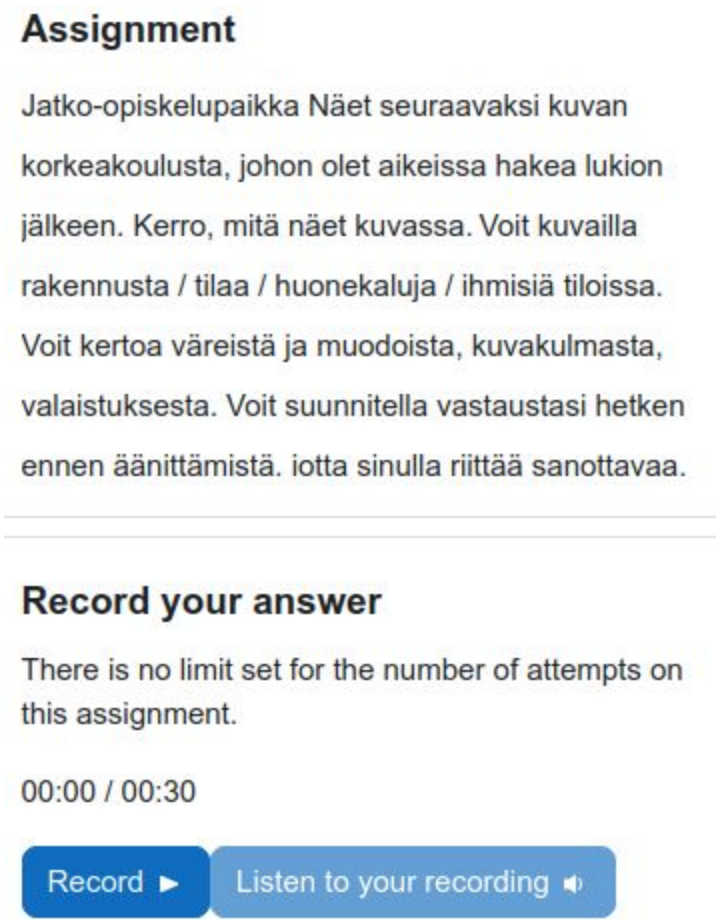
1. Äänite takaisin puhujalle
2. Puheen muunnos tekstiksi (ns. raakatranskriptio)
   a. Ei pisteytetä, mutta näytetään teksti
3. Kunkin äänitteen holistinen taitotasoarvio: A1 - C2 (ennuste)
4. Ääntämisen arvio: 0 - 4p
5. Sujuvuuden arvio: 0 - 4p
6. Kielioppi ja sanasto: 0 – 3p
7. Tehtävän onnistuminen: 0 - 3p



1. Puhumistehtävät

- Luettu puhe
- Vapaa puhe
- Lyhyitä tehtäviä (30 – 60s äänitteet)
- Tilannereagointeja
- Kuvien kuvailua
- Simuloituja haastatteluita
- Eri taitotasoille eri tehtäviä





2. Opetusdatan kerääminen

- Puheiden keräämiseksi luotiin “kokeita” joissa samantyyppisiä tehtäviä kuin on tarkoitus automaattisesti arvostella
- Kerättiin aineistoa lukioissa, yliopistoissa, Yki
- 19 tuntia suomea (325 puhujaa)
- 7 tuntia ruotsia (181 puhujaa)

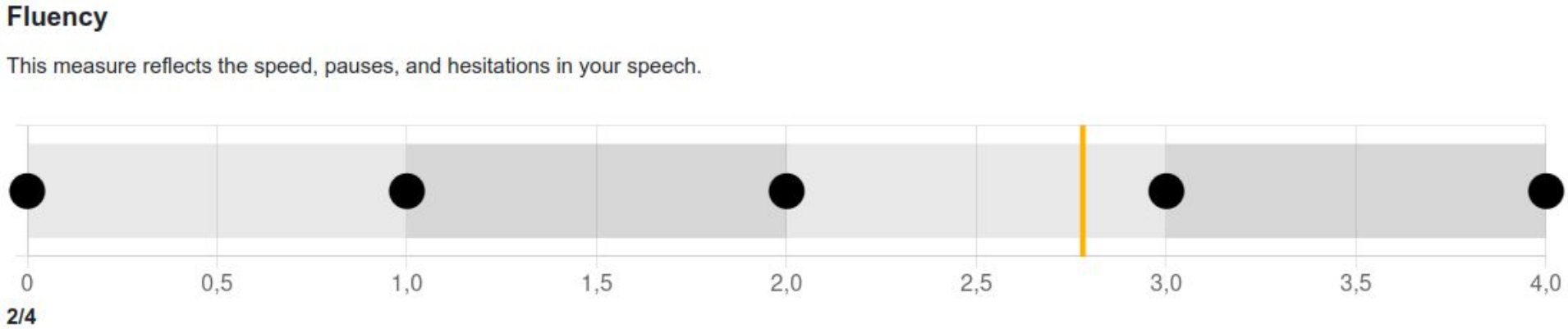
3. Opetusdatan litterointi

- Kaikki kerätty puhe kirjoitettiin auki tekstiksi, myös epäröinnit ja virheet (niinkuin ne kuultiin)
- Kirjattiin myös äänitysongelmat
- Useita litteroijia, mutta kaikki puhe litteroitiin vain kertaalleen
- Litterointia käytettiin sekä puheentunnistimen että arvioinnin opetuksessa

vaikka jokainen *kuvva* on hieman erilainen niissä kuitenkin kaikissa näkyy samanlainen rauhallinen teema sanoisin ja eniten minua kiinnostaa kuva ykkönen siinä on erittäin kaunis väripaletti ja muutenkin maisemakuvat ovat erittäin mielenkiintosia mutta ne voivat mennä tylsän puolelle jos niissä on liikaa vihreetä mutta täs on just tosi kauniisti kuvattu aa<hesitate> vesi ja taivas ja kaikki semmonen hieno yhdistelmä kuva kaks on myös eläinrakkauden puolelta öö<hesitate> ihana myös minulle ja kuva kolme on näyttää minulle tosi tavalliselta ei mitenkään erityiseltä kuvalta jonka voisin itsekkin helposti ottaa tietenkin tulisi huonompi laatu mutta ei se ole mitenkään kiinnostava joten josmä varmaan palkitsisin jonkun näistä instagram<name> kuvista tässä kilpailussa niin se olisi varmasti toi kuva ykkönen sillä mä jotenkin tosi paljon pidän siitä kuvasta mutta koira on myös söpö sii<garbage> joo on erittäin söpö paral>

4. Asiantuntijat arvioivat näytteet

- Tallenteille annettiin sekä holistinen yleisarvosana että analyyttiset osa-aluearviot
- Asiantuntijoina sekä Ykin arviointipooli että kielenopettajia
- Osa näytteistä arvioitiin useampaan kertaan arvioinnin laadun mittaamiseksi
- Työlästä puuhaa, joten kaikkea puhetta ei voitu arvioida
- Työkaluna sama Moodle, jolla puhe näytteet kerättiin, mutta nyt arviointipuoli



5A. Puheentunnistus (puheesta tekstiksi)

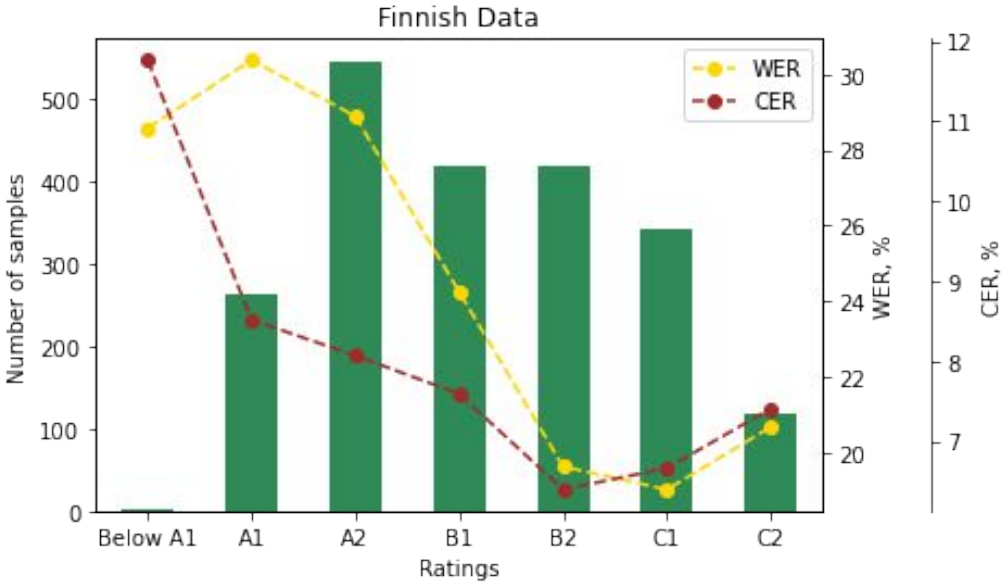
Aalto University

Mallien rakentaminen

- Perustuu **avoimeen** suureen **puhemalliin** jossa noin 100000h puhetta, 30 eri kieltä, ilman litterointia (ruotsilla oma mallinsa)
- Malli ohjattu ja muokattu **suomen puheentunnistukseen** käyttäen Lahjoita Puhetta -aineistoa (litteroitu 1600h)
- Edelleen **muokattu kielenoppijoille** käyttäen DigiTala-aineistoa (16h)
- Ilman kielimallia tunnistaa myös ääntämis- ja muut virheet

Tulokset ja suorituskyky

- Mitä parempi taitotaso sitä tarkempi tunnistus
- Vaikea aineisto => ihmislitterointi ei virheetön



5B. Puheen arviointi (puheesta arvosanoihin)

Aalto University

Mallien rakentaminen

- Perustuu **samaan** muokattuun **puhemalliin** kuin puheentunnistus
- **Muokataan** edelleen kuhunkin **arviointitehtävään** käyttäen DigiTala-aineiston malliarvioita
- Yleisarvosana ei perustu osa-aluearvioihin vaan on oma erillinen mallinsa
- Vertailua ja analysointia varten teimme erilliset mallit puheen tärkeimmille mitattaville piirteille

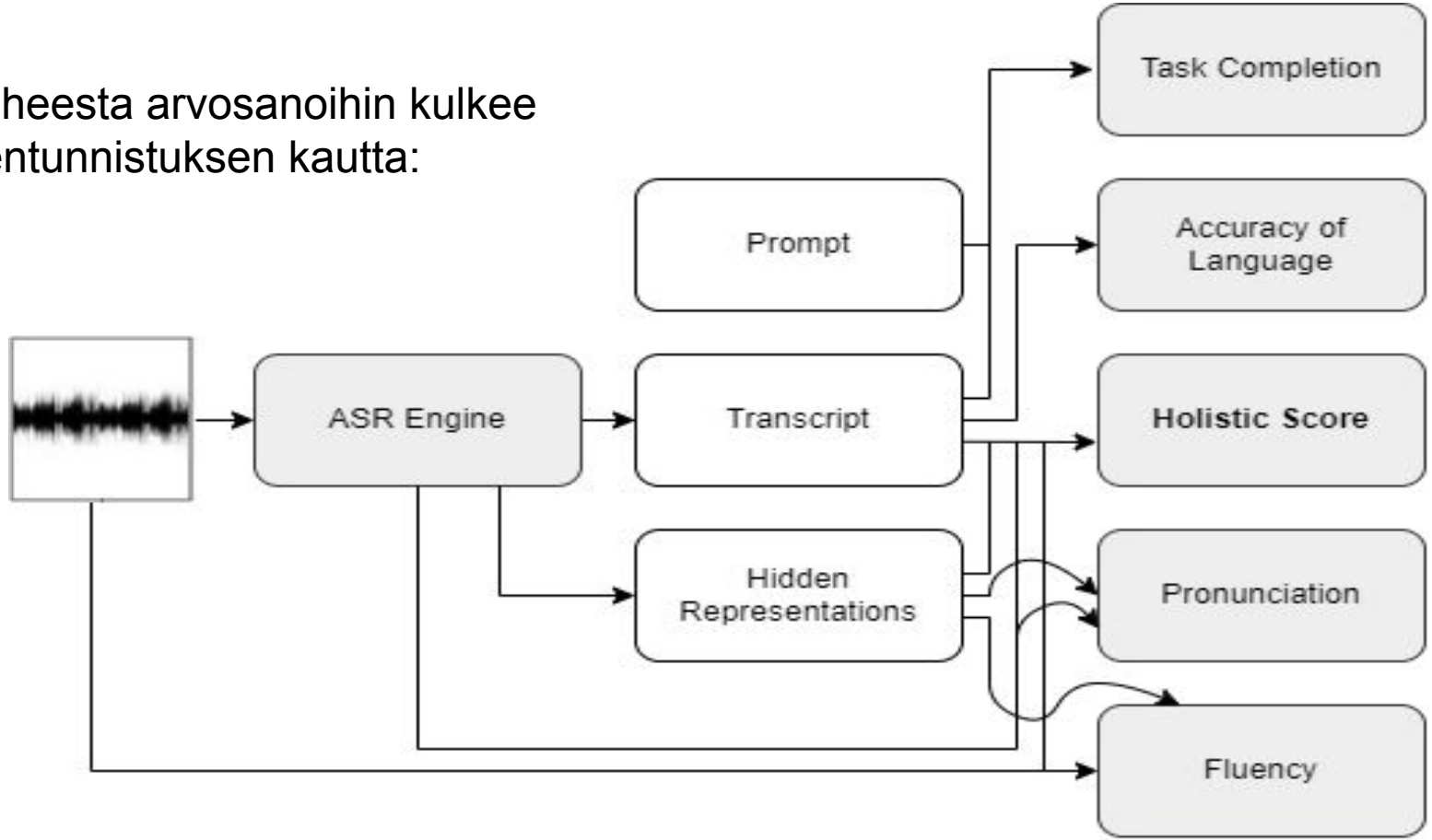
Tulokset ja suorituskyky

- Poikkeaa keskimääräisestä ihmisarviosta vähemmän kuin ihmiset toisistaan
- Human-to-human vs machine-to-human

	Corr. ↑	MAE ↓	Corr ↑	MAE ↓
Holistic(6)	0.75	0.78	**0.80**	**0.61**
Fluency(3)	0.39	0.58	**0.52**	**0.36**
Pronunc.(3)	0.53	0.45	**0.61**	**0.27**
Lex-Gram.(3)	**0.58**	0.40	0.55	**0.27**
TaskAchiev.(3)	0.30	0.41	***0.39***	***0.32***

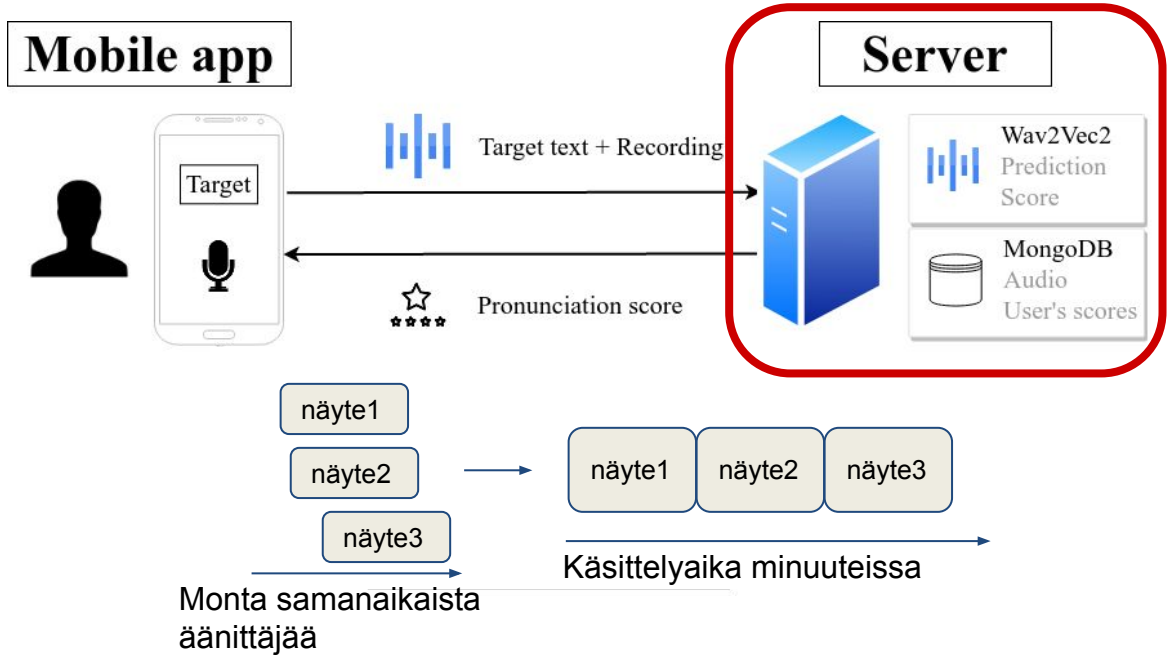
Puheesta arvosanoihin

- Tie puheesta arvosanoihin kulkee puheentunnistuksen kautta:



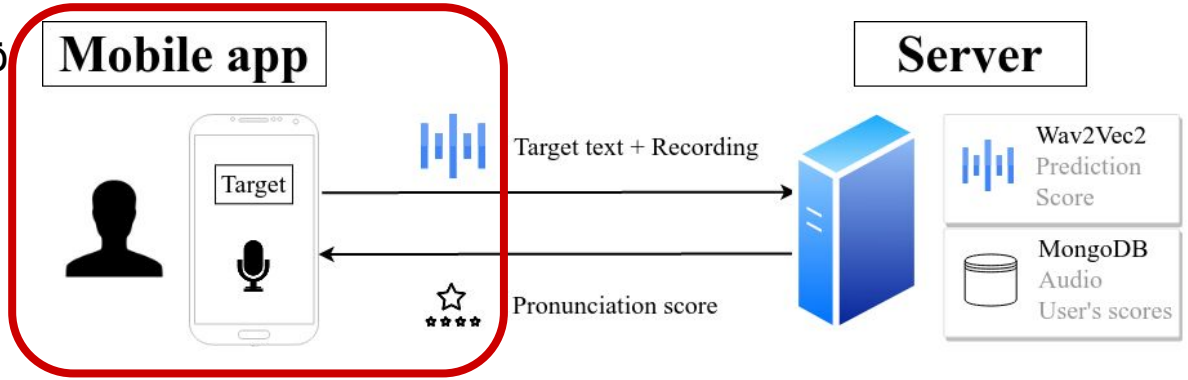
6. Palvelimen rakentaminen

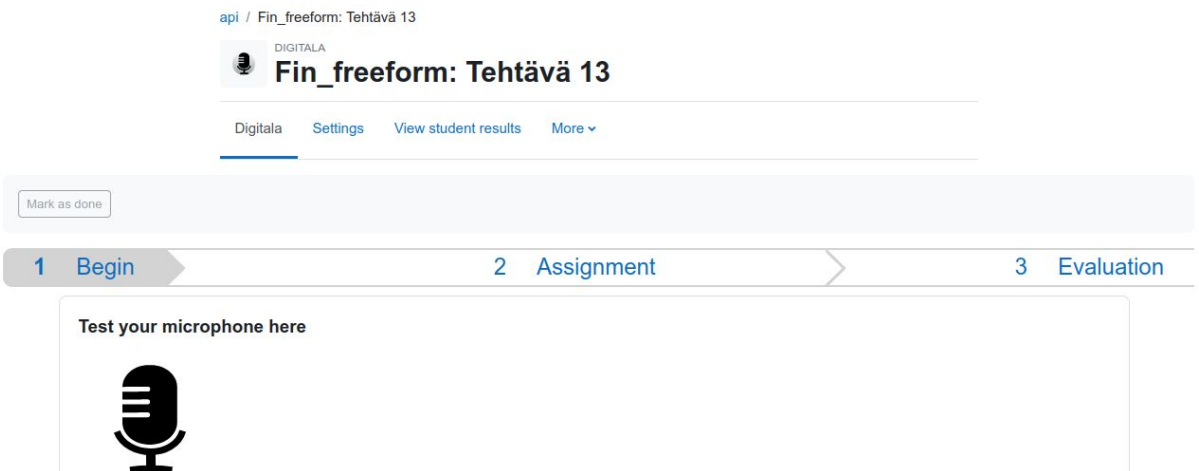
- Vastaanottaa puhenäytteet, tallentaa ja palauttaa tekstit ja arvosanat
- Tunnistusmallit ovat suuria joten laskeminen vaatii aikaa
- Mitä pidempi tallenne sitä enemmän laskettavaa
- Mitä enemmän käyttäjiä sitä pidemmät jonot ja viiveet
- Skaalautuu lisäämällä laskentapalvelimia



7. Käyttöliittymän rakentaminen

- HY:n opiskelijoiden harjoitustyö
- Moodlessa puhetehtävien teko ja parametrien määrittely helppoa
- Moodle hoitaa myös käyttäjien hallinnoinnin ja salasanat
- Käyttöliittymä äänittää ja lähettää puheet palvelimelle
- Kun tulokset saapuvat ne esitetään käyttäjälle





8. Testaus ja palautteen keruu

Stressitestit

- Tutkijat
- Ovatko konearviot järkeviä
- Nopeuden pullonkaulat
- Paljonko palvelin hidastuu kun käyttäjämäärä lisääntyy
- Kuinka monta laskentapalvelinta tarvitaan tietylle määrälle käyttäjiä

Käyttäjäpalaute

- Kielenopiskelijat ja -opettajat
- Opettajat voivat tarkastaa omien opiskelijoidensa saamat arviot
- Suhtautuminen koneelliseen arviointiin
- Arvioinnin luotettavuus ja palautteen hyödyllisyys
- Kehitysideat

Aalto University

Kiitos.

Questions that we have tried to answer in DigiTala

1. How to do Automatic Speaking Assessment (ASA) using Automatic Speech Recognition (ASR)?
2. Do self-supervised learning and pre-trained models solve the ASR of spontaneous speech in under-resourced L2 languages?
3. Which features to use for computing ASA? Explainable ones vs black box?
4. Which of these features have most effect in human speaking assessment?
5. How to make a classifier to get the overall proficiency level from the features?
6. Can we predict the overall level from the assessments in the analytic dimensions (aspects) of speaking skills?
7. Do the ASR (speech-to-text) errors matter for ASA?
8. How to compute feedback using the various aspects of speaking skills?

Answers, so far

1. Al-Ghezi, Getman, Rouhe, Hildén and Kurimo. Self-supervised End-to-End ASR for Low Resource L2 Swedish. Proc. **Interspeech 2021**
2. Yaroslav Getman. End-to-End Low-Resource Automatic Speech Recognition for Second Language Learners. MSc thesis. Aalto University, October 2021. **“Yaroslav”**
3. Al-Ghezi, Voskoboinik, Getman, von Zansen, Kallio, Akiki, Kuronen, Kurimo, Huhta and Hildén. Automatic Speaking Assessment of Spontaneous L2 Finnish and Swedish. (in review) **“LAQ”**
4. Al-Ghezi, Getman, Singh and Kurimo. Automatic Rating of Spontaneous Speech for Low-Resource Languages. (in review) **“SLT”**
5. Clara Akiki. Automatic Assessment of Spoken Lexico-Grammatical Proficiency in L2 Finnish and Swedish. MSc thesis. Aalto University, August 2022. **“Clara”**

DigiTala

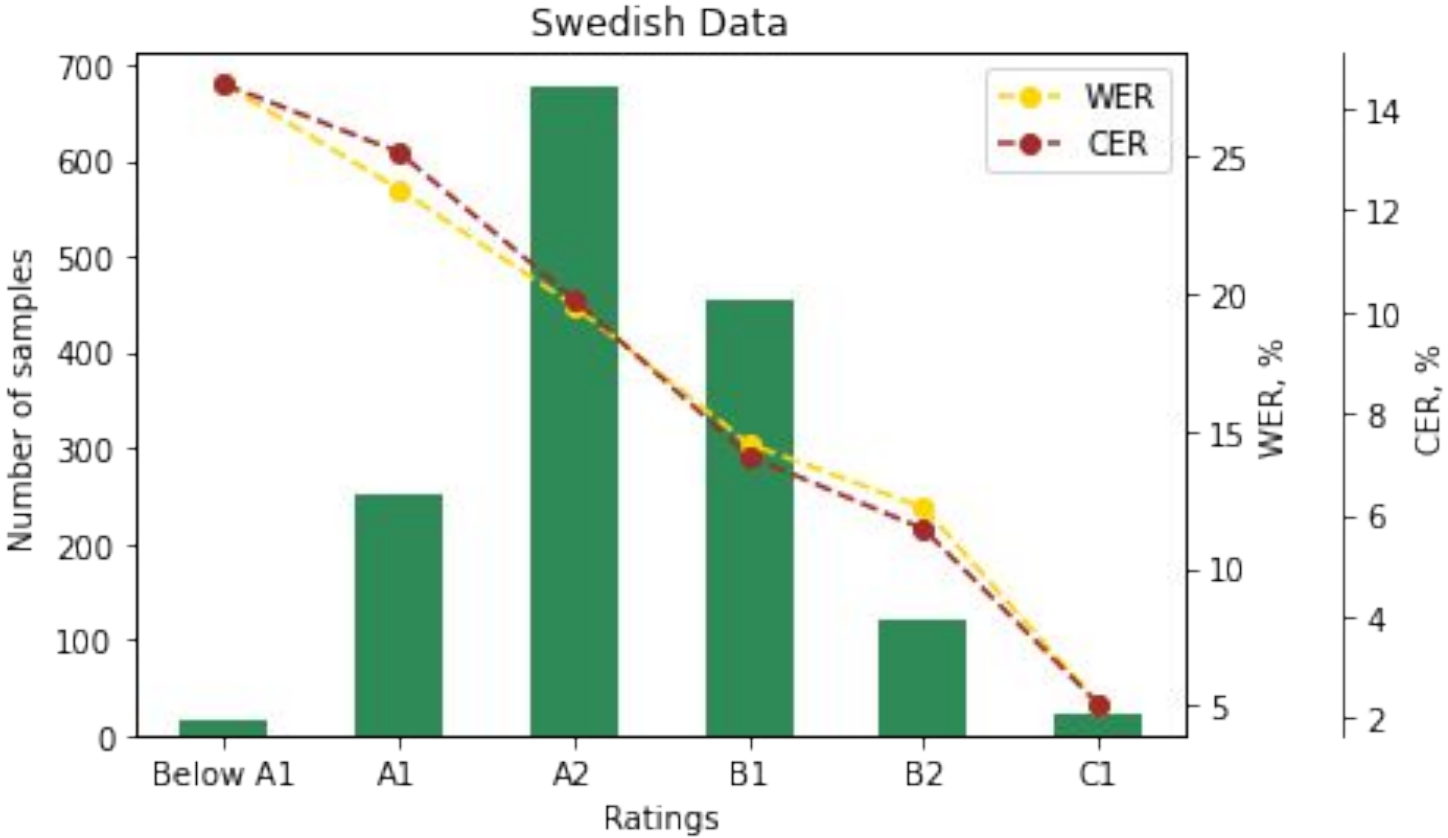
DigiTala

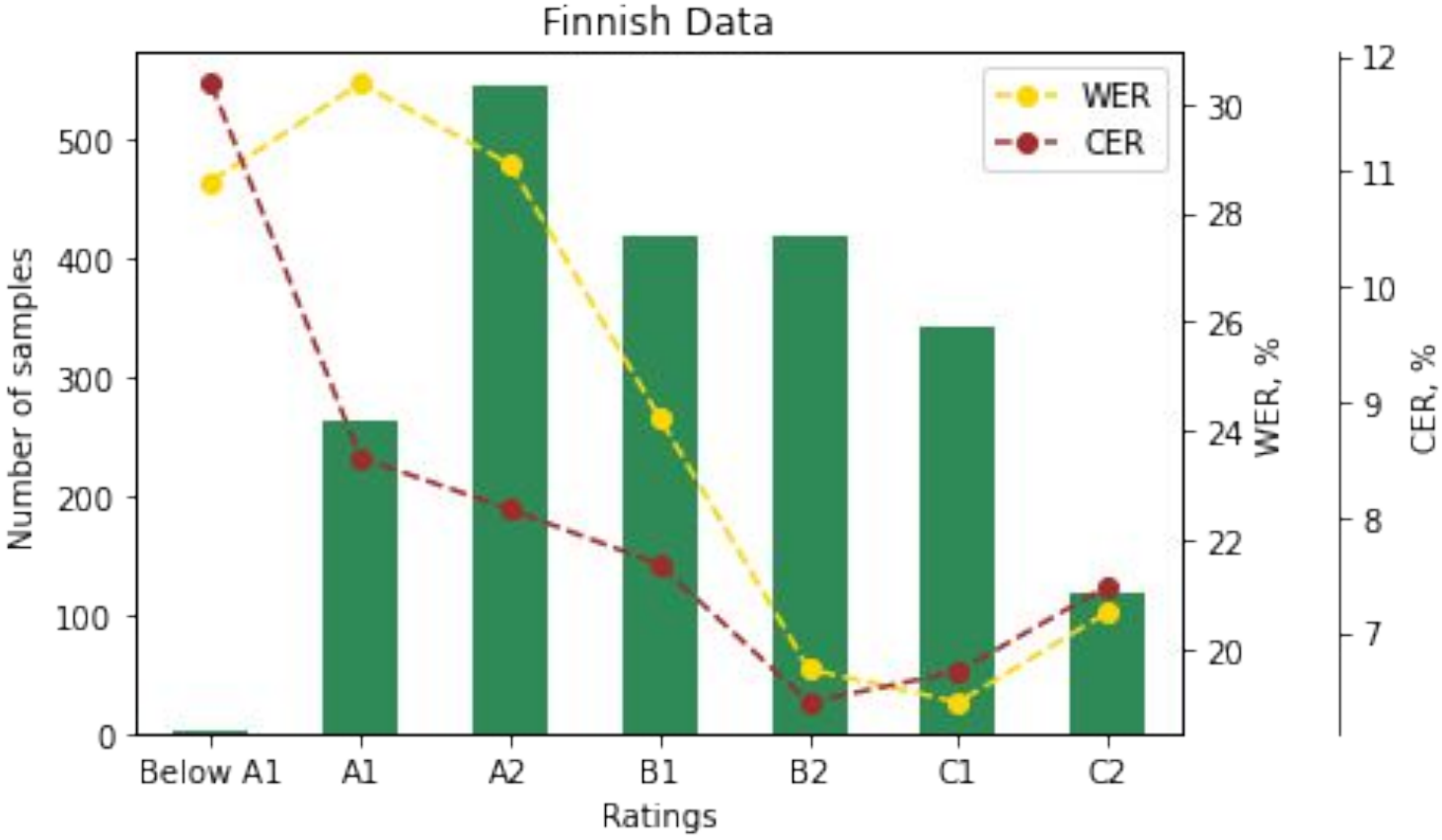
L2 speech data for ASA experiments

- Swedish (178) and Finnish (308) L2 speakers
- Responses to tasks such as describe a picture or reply to a question
- Swedish: 1547 samples, 5.6 hours
- Finnish: 2112 samples, 14.1 hours

- Samples transcribed and rated by human experts in 7-level CEFR-related scale (A1 - C2)
- Swedish rated from 2 - 5 (mostly 3)
- Finnish rated from 2 - 7 (mostly 3 - 6)
- Wav2Vec2 based ASR with 17.3% WER in Swedish and 16.2% in Finnish

[*Al-Ghezi, Voskoboinik, Getman, von Zansen, Kallio, Akiki, Kuronen, Kurimo, Huhta and Hildén. Automatic Speaking Assessment of Spontaneous L2 Finnish and Swedish. (in review)]*


Swedish Data
WER
CER
Number of samples
WER, %
CER, %
Ratings
Below A1
A1
A2
B1
B2
C1
0
100
200
300
400
500
600
700
5
10
15
20
25
2
4
6
8
10
12
14


Finnish Data
WER
CER
Number of samples
WER, %
CER, %
Ratings
Below A1
A1
A2
B1
B2
C1
C2
0
100
200
300
400
500
20
22
24
26
28
30
7
8
9
10
11
12

DigiTala

PFL vs Wav2Vec2 features for overall grading

Both features are combined with 6-layer DNN classifier and measured vs human rating in Swedish and Finnish:

- Measures: Precision, Recall, F1, Spearman corr.
- Selected PFL features:
  - Pronunciation, Fluency, Lexical
- Wav2Vec2 features (1024-dim, layer 12/24)
  - Pre-trained with unsupervised L1 speech
  - Fine-tuned for L1 ASR (W2V2)
  - + fine-tuned for ASA (W2V2 CE)

[Al-Ghezi, Getman, Singh and Kurimo. Automatic Rating of Spontaneous Speech for Low-Resource Languages. (in review)]

Swedish	F1	Spearman
PFL	0.41	0.42
W2V2	0.48	0.54
PFL+W2V2	0.50	0.56
W2V2 CE	**0.56**	**0.64**

Finnish	F1	Spearman
PFL	0.37	0.75
W2V2	**0.41**	0.80
PFL+W2V2	0.39	**0.82**
W2V2 CE	0.39	0.80

DigiTala PFL features that affect most for the overall grade

Pronunciation

- Acoustic model score
- Voiced fraction
- Average syllable distance
- Consonant and Vowel Duration
- Consonant and Vowel Logits

Fluency

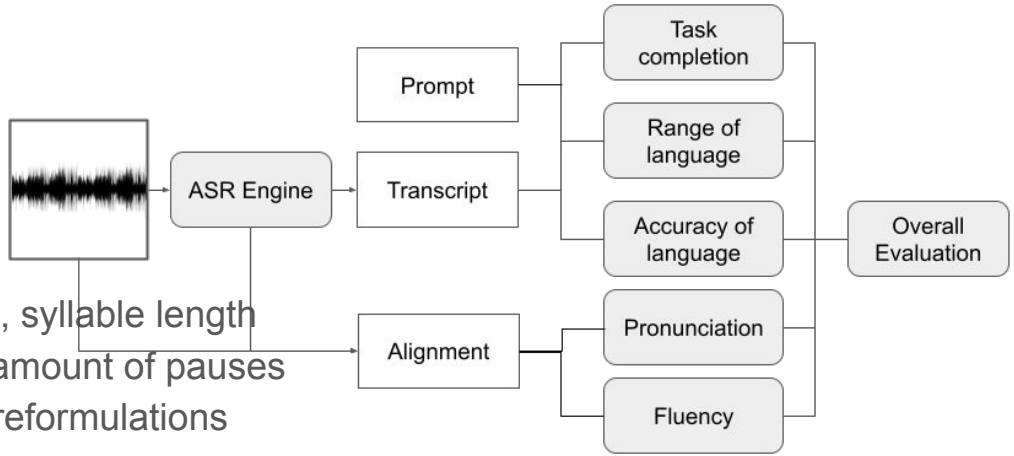
- Speech rate:
- Articulation rate
- Pairwise Vocalic Indexing (rPVI cons)

Lexical

- Length in words
- types/tokens ratio
- #tokens

Feedback? - 5 main aspects (dimensions) in human assessment

- **Pronunciation**:
  - individual sounds
  - prosody (intonation and stress)
- **Fluency**:
  - speed: speech and articulation rate, syllable length
  - breakdown: frequency, length and amount of pauses
  - repair: repetitions, false starts and reformulations



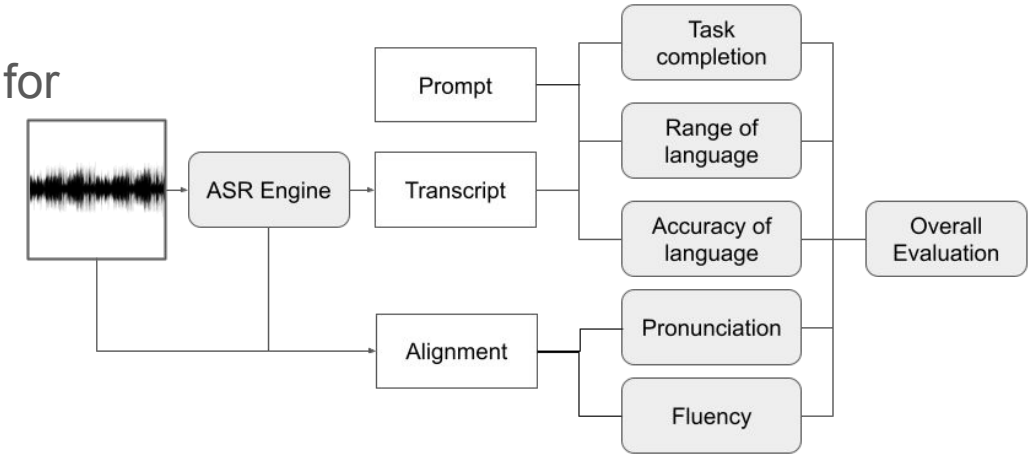
- **Range**: Lexical diversity + grammatical complexity
- **Accuracy**: Impact of lexical + grammatical errors on comprehensibility
- **Task achievement**: The correct topic/question/answer
- In addition to the Overall Grade, the human raters were asked to rate each sample in scale 1 - 4 for each of these 5 main dimensions

DigiTala

[*Al-Ghezi, Voskoboinik, Getman, von Zansen, Kallio, Akiki, Kuronen, Kurimo, Huhta and Hildén. Automatic Speaking Assessment of Spontaneous L2 Finnish and Swedish. (in review)*]

To compute feedback? Train 5 Evaluators to estimate analytic scores

- Train the automatic Evaluators for
- **Fluency**: Acoustic/prosodic features computed from the aligned audio signal
- **Pronunciation**: ASR-based features derived using the acoustic models



- **Range and Accuracy:** Lexical features computed from the recognition output and grammatical features obtained using parsers
- **Task achievement**: Computing the semantic distance to the correct answers (or prompt). Filter out prompt repetitions and other invalid answers.

DigiTala

[*Al-Ghezi, Voskoboinik, Getman, von Zansen, Kallio, Akiki, Kuronen, Kurimo, Huhta and Hildén. Automatic Speaking Assessment of Spontaneous L2 Finnish and Swedish. (in review)]*

Performance of analytic Evaluators and holistic grade

- Decision Tree (DT) classifiers 5-fold cross validation, DNN for Overall
- Measured vs human ratings by F1 (precision+recall) and Spearman corr.
- **Fluency**: 3 levels, rates, pauses
- **Pronunciation**: 2 levels, AM score, min.pitch, vowel & cons.durations
- **Range**: 2 levels, lexical profile
- **Accuracy**: 3 levels, HumanSR vs ASR
- **Overall** Fin: 5 classes (Human ≈0.32A +0.06P +0.71R +0.49F +0.27T)
- **Overall** Swe: 3 classes (Human ≈0.27A +0.12P +0.45R +0.25F +0.44T)

	Fin-F1	Fin-S	Swe-F1	Swe-S
Fluency	0.55	0.46	0.39	0.23
Pronunciation	0.33	0.12	0.58	0.17
Range HSR	**0.61**	**0.28**	**0.55**	**0.20**
Range ASR	0.57	0.18	0.37	0.13
Accuracy HSR	0.42	0.22	0.35	**0.18**
Accuracy ASR	0.42	**0.23**	**0.37**	0.13
Overall DT	0.34	0.32	0.33	**0.14**
Overall DNN	**0.60**	**0.55**	0.33	0.12

[*Al-Ghezi, Voskoboinik, Getman, von Zansen, Kallio, Akiki, Kuronen, Kurimo, Huhta and Hildén. Automatic Speaking Assessment of Spontaneous L2 Finnish and Swedish. (in review)*]

DigiTala

ASA for lexicon and grammar using ASR transcripts

- S = Spearman corr. E =mean squared error
- Not significantly worse (<10%) than human transcripts
- ASR affects more accuracy
- Hand-crafted features comparable to BERT
- BERT features: Roberta slightly better than BERT
- Scoring: Random Forest RF comparable to DNN

	Swe-S	Swe-E	Fin-S	Fin-E
Range-BERT-RF	**0.44**	**0.12**	0.21	0.15
Range-HAND-RF	**0.44**	**0.12**	0.18	0.15
Range-BERT-DNN	0.39	0.13	**0.28**	**0.14**
Range-HAND-DNN	0.39	0.13	0.25	0.15
Accuracy-BERT-RF	**0.30**	**0.22**	**0.36**	0.30
Accuracy-HAND-RF	**0.30**	**0.22**	0.35	0.30
Accuracy-BERT-DNN	0.23	0.23	0.33	0.30
Accuracy-HAND-DNN	0.23	0.24	0.30	0.33

[Clara Akiki. Automatic Assessment of Spoken Lexico-Grammatical Proficiency in L2 Finnish and Swedish. MSc thesis. Aalto University, August 2022.]

DigiTala

DigiTala

Conclusions, so far

- ASR works reasonably well (for ASA) despite the lack of training data
- Explainable features work ok, but black box features usually better
- Explainable classifiers work ok, but black box classifiers usually better
- The scores in analytical dimensions provide more detailed feedback, although their effect in the overall grade is harder to explain
- The unbalanced classes and inter-annotator disagreement are challenging
- Project video Demo Zenodo

For more information

- Contact: *mikko.kurimo@aalto.fi*
- Publications: http://research.aalto.fi
- Home page: (search: *”Aalto asr home”)*
- Software: (search: *”Aalto asr github”)*
- Demos: (search: *”Aalto asr video”)*