Classifying if a person knows a word using speech data

Marijn Schraagen

NLP Group Meeting

October 26th, 2023

Universiteit Utrecht

Project overview

- AI Labs project about low literacy in Dutch high school students
- Collaboration between Computer Science (me, Mehdi Dastani), Social Science (Hans Marien, Henk Aarts), Humanities (Els Stronks)
- Goal to start an ELSA Lab (Ethical, Legal, Societal Aspects) in NL AI Coalition
- Topics: speech and text models for low literacy, (personalized) interventions
- Today: speech classification for **vocabulary knowledge**

Data

- Data collection for a separate FSS master project
- Data re-used by me for different experiments
- *Discussion:* can we collect more suitable data?
- Recordings of 40 single Dutch words repeated from TTS spoken prompts
- 50 high-literate participants, mostly students, native speakers
- Questionnaire at the end with many questions
- Did you know the words that are used in the experiment?

```
keen  - yes no
rente - yes no
twee  - yes no
zijp  - yes no
omzet - yes no
zeeg  - yes no
```

Research goal

Assumption: if a person knows a word then their voice will sound different – confidence, fluency, pitch, lower-level features

Assumption: this is measurable in principle

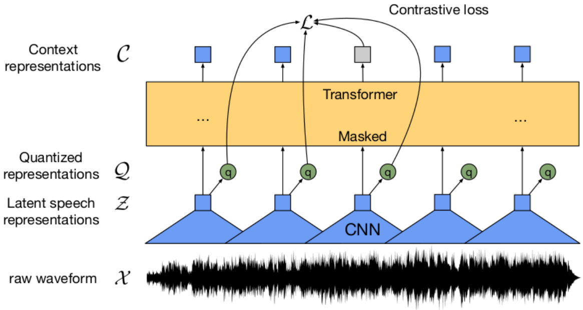
Assumption: current (transformer) models for audio can pick up on this

Assumption: the collected data exhibits the characteristic under study

Research goal: Build a model to classify from a spoken word if the person knows this word

Experiments

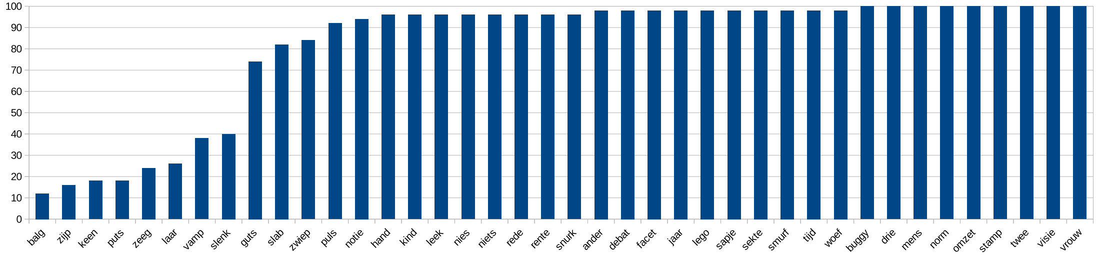
Classification task using the `wav2vec` model with a classification head for the binary *word_known* variable



Universiteit Utrecht

Data imbalance

- Data is imbalanced: 82% known word examples
- Use all 734 unknown word examples and an equal number of random known word examples
- Words range from largely unknown to known by everyone (more on that later)



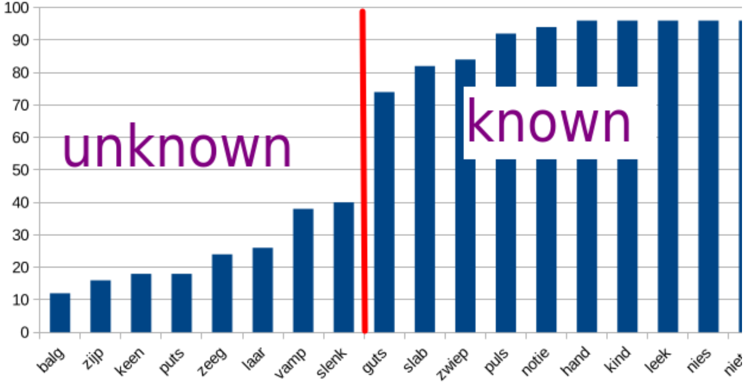
Results

```
              precision    recall  f1-score   support

     unknown       0.82      0.88      0.85       110
       known       0.87      0.81      0.84       111

    accuracy                           0.85       221
```

Not bad?

Baseline

- Random baseline: 50%
- Second baseline: majority label *per word* from the training set



Results majority baseline

Trained classifier:

```
              precision    recall  f1-score   support

     unknown       0.82      0.88      0.85       110
       known       0.87      0.81      0.84       111

    accuracy                           0.85       221
```

Majority baseline:

```
     unknown       0.82      0.84      0.83       111
       known       0.83      0.81      0.82       110

    accuracy                           0.82       221
```

Bad! The model just learns to predict the word.

Word split

- Rearrange train and test: split on vocabulary items
- Six words manually selected for test set: from frequently unknown to frequently known, other 34 words in training set

```
              precision    recall  f1-score   support

     unknown       0.49      0.13      0.20       188
       known       0.33      0.76      0.46       104

    accuracy                           0.35       292
```

Very bad!

Adding information

- Hypothesis: if a speech recognition system has trouble recognizing the word then the person did something strange – and may not know the word
- Far fetched? Let's try :)
- Three models: Kaldi-NL, Whisper small, Whisper large (with `language="nl"`)
- Whisper models are highly accurate, but regularly hallucinate

word	transcription
notie	mankind. Uwanners van vandaag van de VLaughter. Bedankt voor uw tijd.
debat	대박 (Korean, pronunciation: *daebak*)
vamp	FAM vieleLijke
ander	Am there. Sorry voor het idee! Alująsfteroki
stomp	Stomp Stomp Stomp Stomp Stomp Stomp Stomp
notie	然后勾她一点

- Kaldi doesn't hallucinate and makes more mistakes (=good!)

Adding speech recognition

- Perform ASR on the recording
- Input for the model: recording (audio) + ASR transcription (text)
- Multimodality for the lazy: use TTS to convert transcription back to audio and concatenate

```
             precision    recall  f1-score   support

           0       0.95      0.88      0.92       111
           1       0.89      0.95      0.92       110

    accuracy                           0.92       221
```

Seems better...

Adding speech recognition

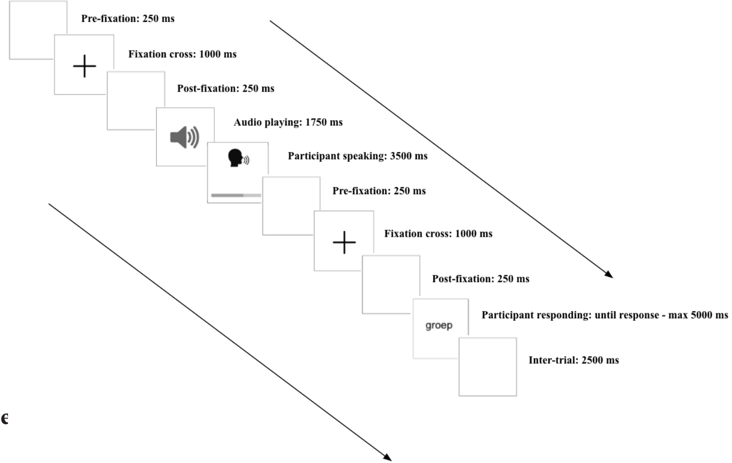
Audio + transcription on word split:

```
              precision    recall  f1-score   support

           0       0.37      0.16      0.23       188
           1       0.25      0.50      0.33       104

    accuracy                           0.28       292
```

Still bad

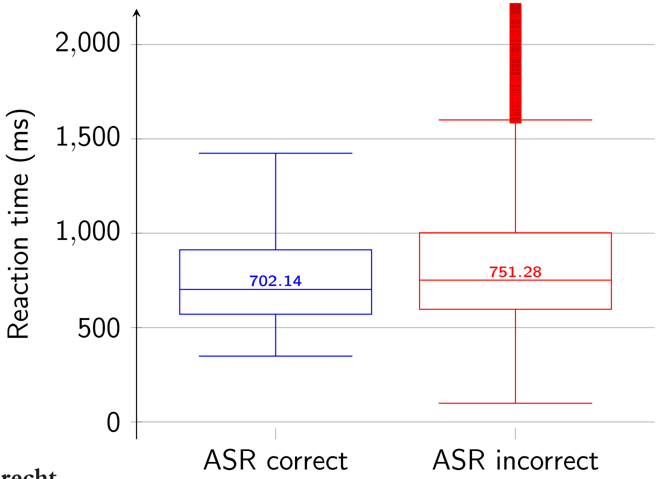
Relation ASR and words known

Experimental task: after reading a word participants were shown either that word (A) or a different word (B) and has to press button A or button B



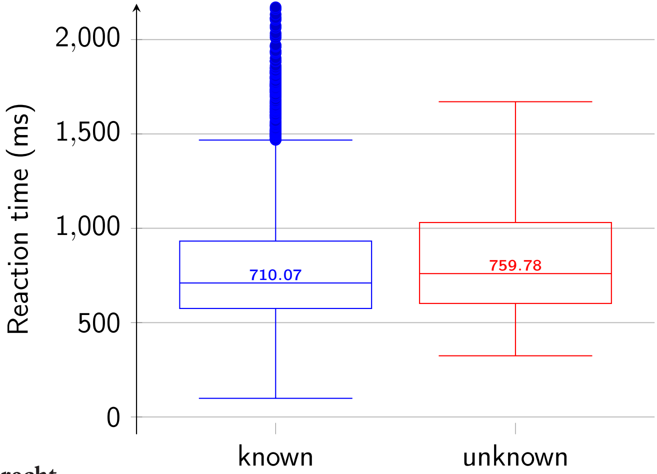
Universite

Reaction time vs ASR correct



Universiteit Utrecht

Reaction time vs word known



Universiteit Utrecht

ASR correct vs. known–unknown

	known	unknown
ASR correct	2096	329
ASR incorrect	1170	405

Preliminary conclusions

- The content of the word is rather important
    - Models like Word2Vec are trained on content
- ASR performance does seem to contain information
    - Even if you don't know if the ASR was right!
- Speaking style by itself does not seem to be sufficient

Issues

- Speech may not carry this kind of information
- Models are not pretrained to support this task
- The dataset is too small, the words too short, the participants too literate
- Asking people to self-report if they know a word is vague and unreliable
- Repeating a spoken word is easier than reading it from a screen or paper

Next steps: data collection

- Data issue can be addressed
- Reduce the influence of content
  1. Speak an unknown word
  2. Learn this word: see/create example sentences, answer questions etc
  3. Speak the same word again
  4. Train models to find the difference between the two recordings

Next steps: data collection

- Data issue can be addressed
  - Read a word instead of repeating a spoken word
  - Test vocabulary knowledge by asking to provide a definition
  - Speak a longer sentence, post-process to isolate the word
  - Use longer words
  - Ask low-literate participants (but: first more trials)

  - Use actual multimodal model
- Suggestions?