SPADE

SPeech Across Dialects of English

Automatic classification of stop realisation with wav2vec2.0

James Tanner, Morgan Sonderegger, Jane Stuart-Smith, Jeff Mielke, Tyler Kendall

Interspeech 2025, Rotterdam

17-21st August 2025

University of Glasgow

McGill

NC STATE UNIVERSITY

Duke UNIVERSITY

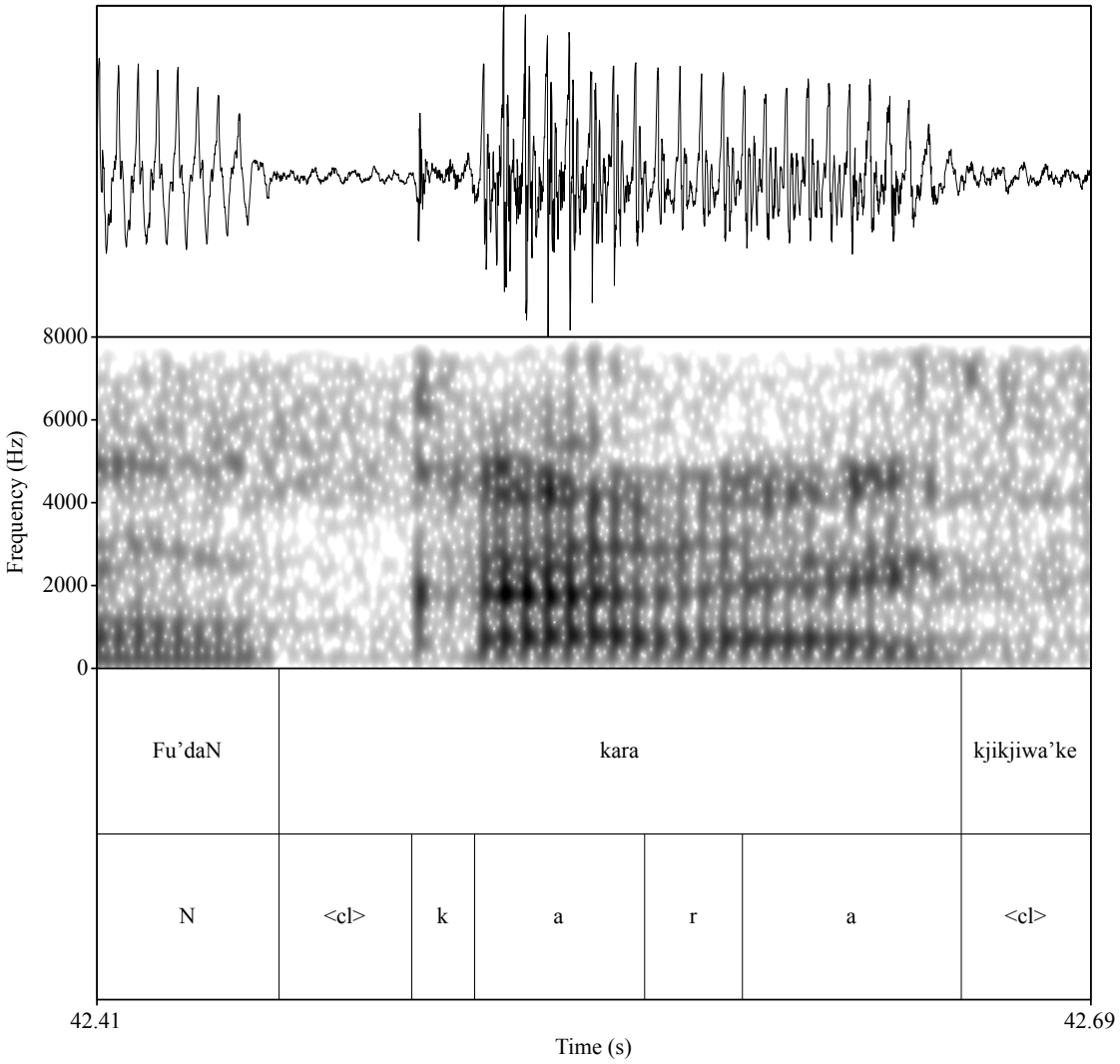
Introduction

- Tools for the automatic/semi-automatic annotation and measurement of speech data essential in modern phonetic research
  - Force-alignment (MFA, P2FA, FAVE, …)
  - Formant tracking (FastTrack, FAVE, PolyglotDB, …)
  - Stop Voice Onset Time (AutoVOT, DrVOT)
- Datasets in phonetic research becoming increasingly large and variable
  - Important that these tools are both accurate and robust to variability

Yuan & Liberman (2008), Keshet & Sonderegger (2014), Rosenfelder et al. (2014), McAuliffe et al. (2017)

Introduction

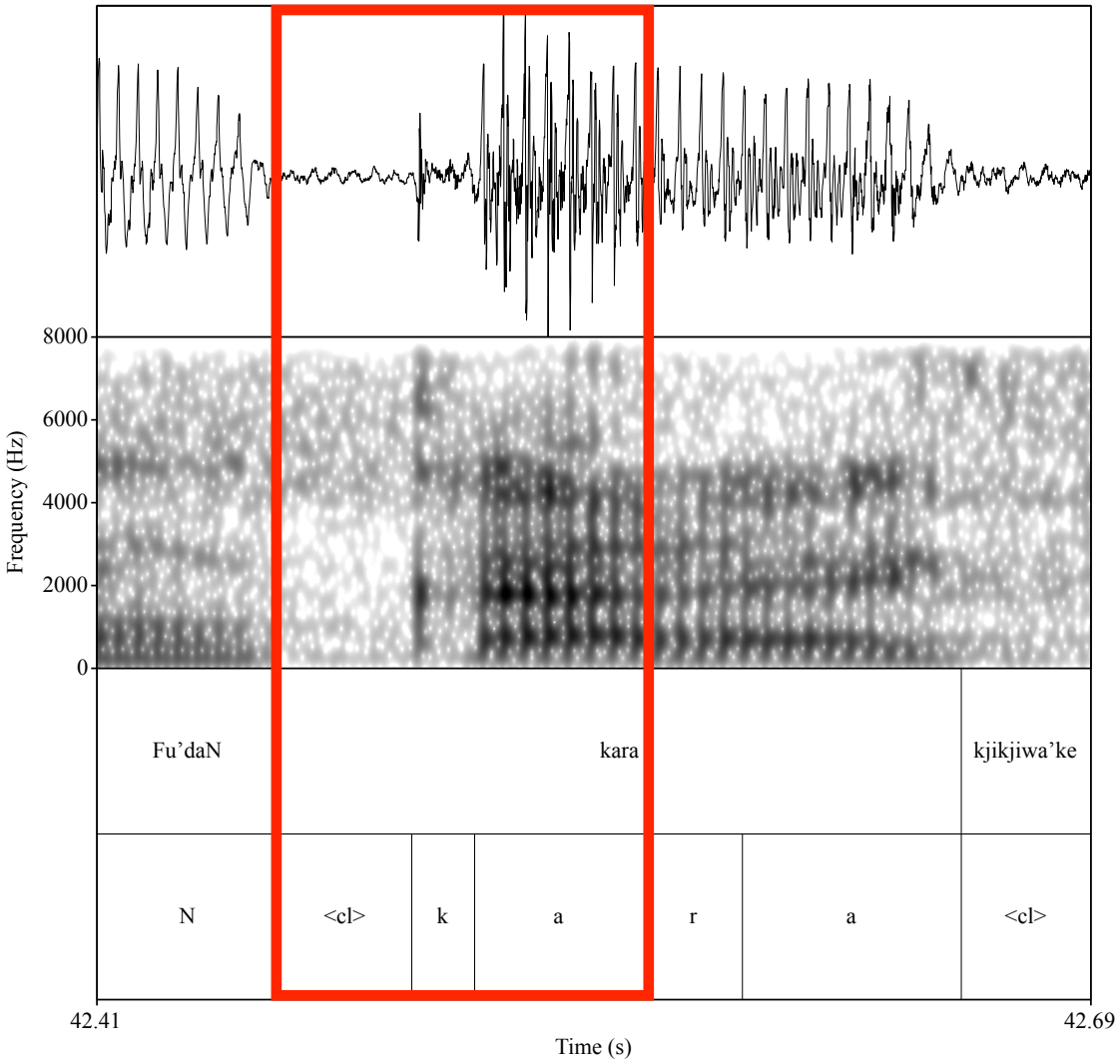
- One type of variability: presence or absence of stop closure + burst



Lavoie (2001), Ladd & Scobbie (2003), Bouavichith & Davidson (2013), Katz & Fricke (2018)

Introduction

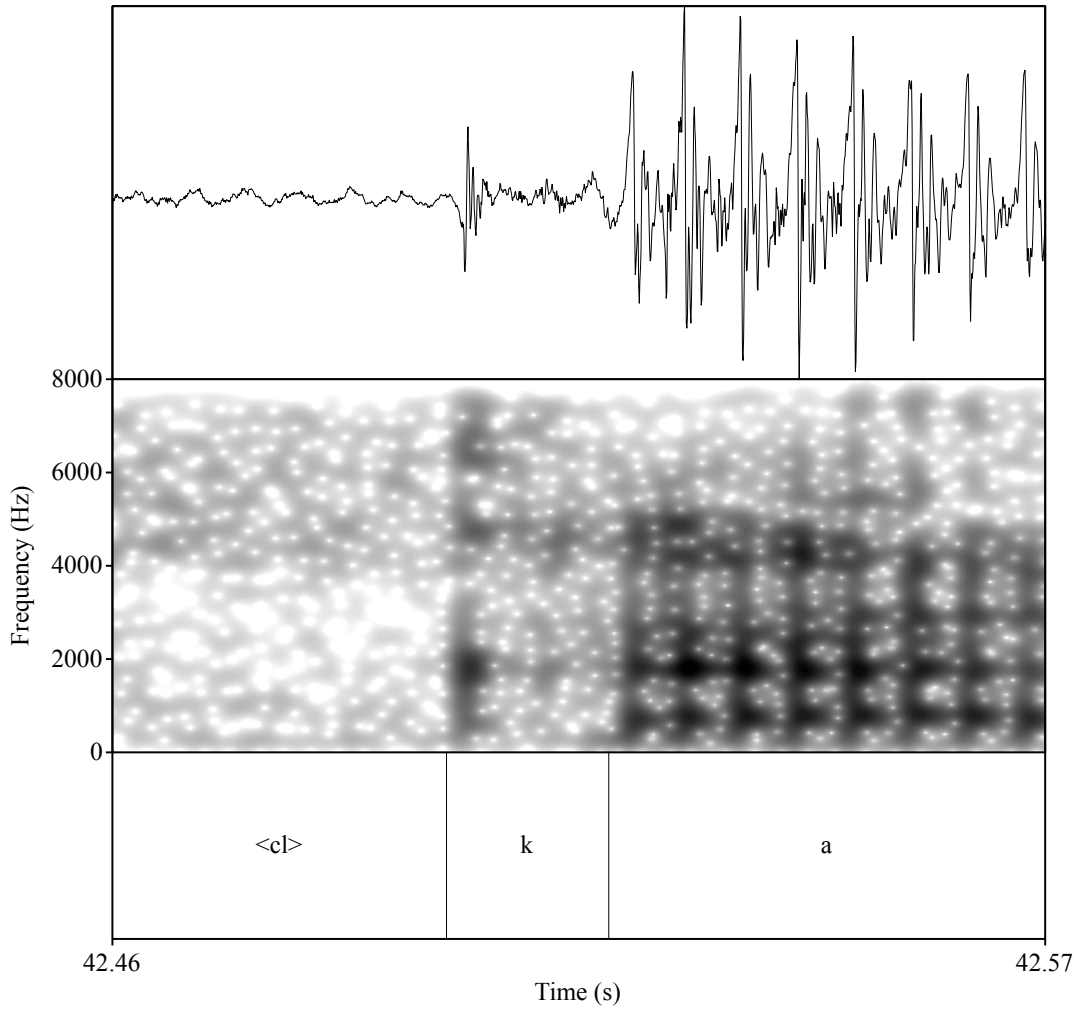
- One type of variability: presence or absence of stop closure + burst



Lavoie (2001), Ladd & Scobbie (2003), Bouavichith & Davidson (2013), Katz & Fricke (2018)

Introduction

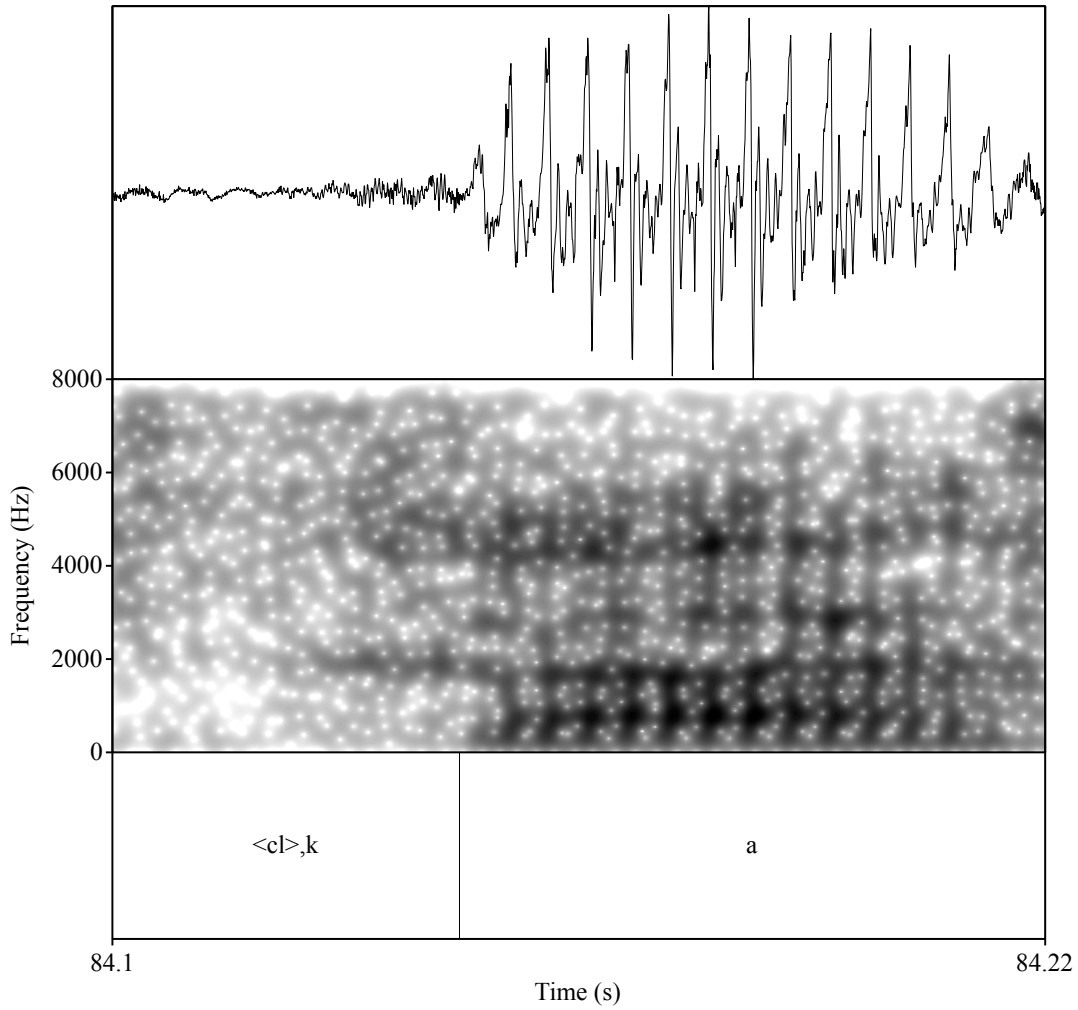
- One type of variability: presence or absence of stop closure + burst
- Stops are realised in a range of ways in natural speech:
  - Full closure + burst



Lavoie (2001), Ladd & Scobbie (2003), Bouavichith & Davidson (2013), Katz & Fricke (2018)

Introduction

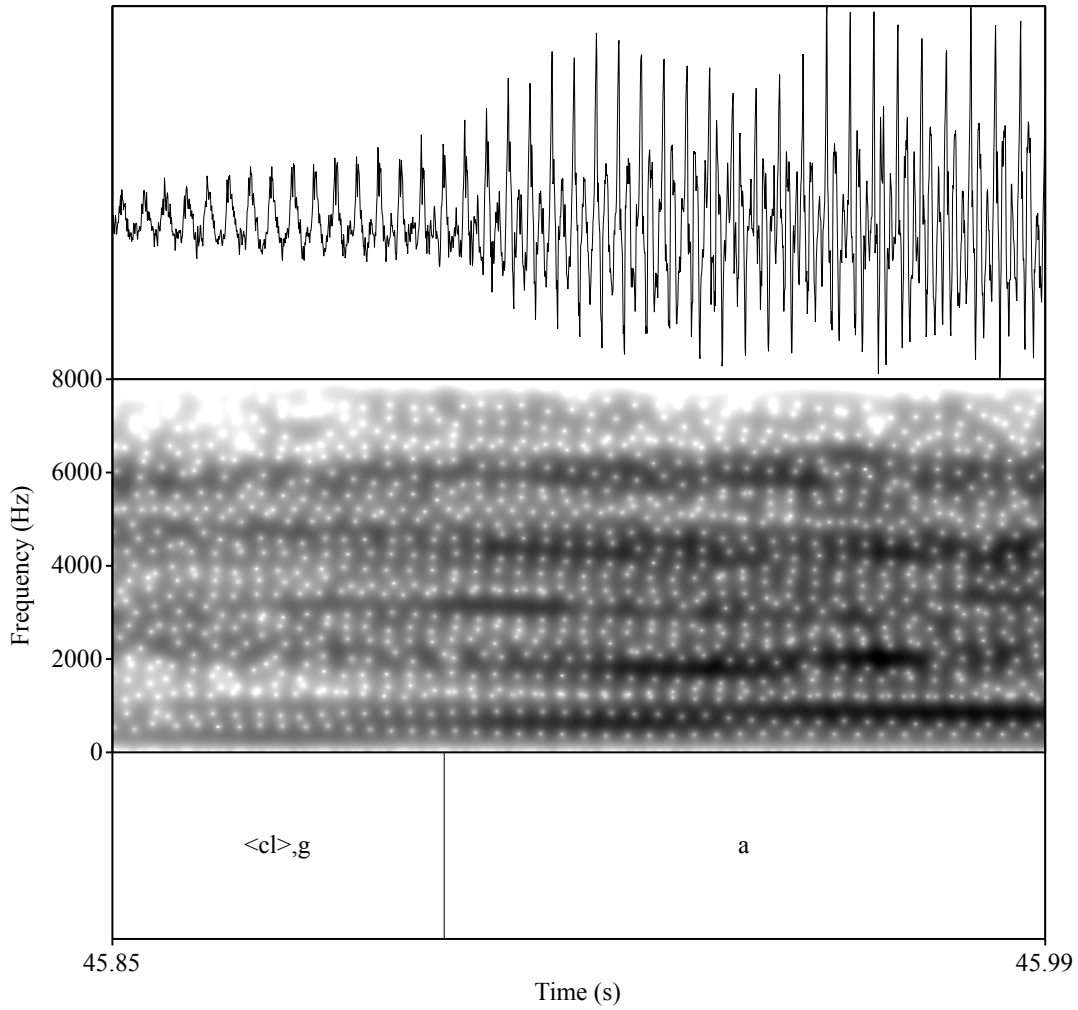
- One type of variability: presence or absence of stop closure + burst
- Stops are realised in a range of ways in natural speech:
  - Full closure + burst
  - Fricative/approximant



Lavoie (2001), Ladd & Scobbie (2003), Bouavichith & Davidson (2013), Katz & Fricke (2018)

Introduction

- One type of variability: presence or absence of stop closure + burst
- Stops are realised in a range of ways in natural speech:
  - Full closure + burst
  - Fricative/approximant
  - Incomplete closure



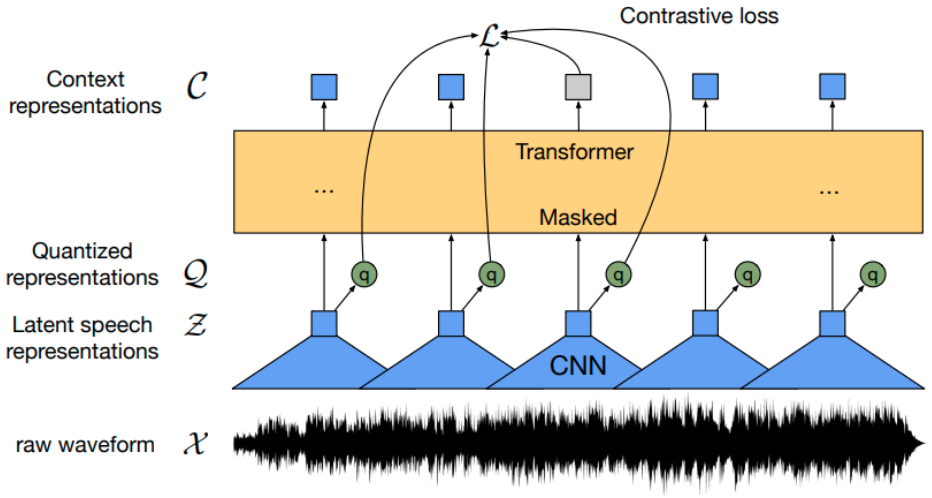
Lavoie (2001), Ladd & Scobbie (2003), Bouavichith & Davidson (2013), Katz & Fricke (2018)

Introduction

- Information of stop realisation within a corpus/dataset useful for:
  - Study of lenition/stop realisation patterns in their own right;
  - Downstream quantitative measurement (i.e. VOT, burst amplitude, etc.)
- No widely-available tool for automatically annotating stop realisation in a speech corpus
- Here: present a method & tool for predicting the realisation of stops using *wav2vec2*

Introduction: wav2vec2

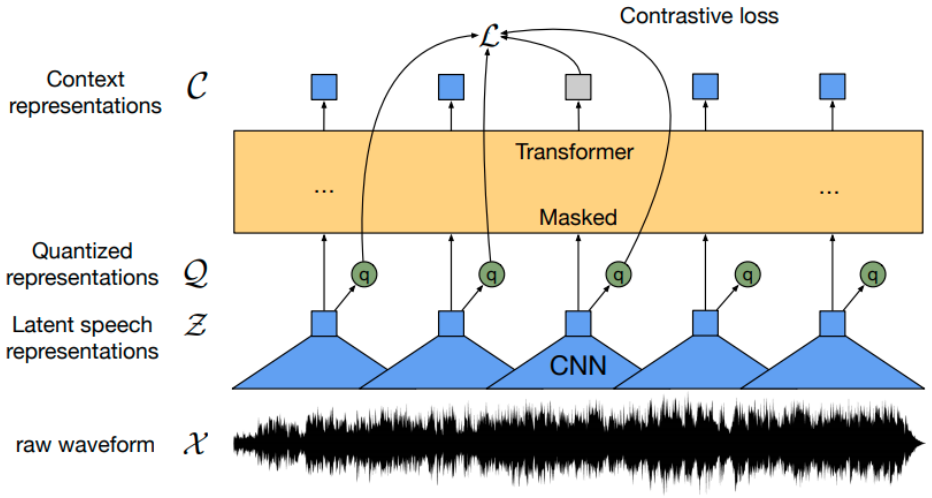
- Convolutional neural network (feature encoder) + transformer
- Pre-trained on large amounts of unlabeled speech data
  - Next time-step prediction
  - Masked time-step prediction
- Can then fine-tuned for a specific task (speech recognition, emotion classification)



Schneider et al. (2019), Baevski et al. (2020)

Introduction: wav2vec2

- Shown to be highly sensitive to phonetic information in both training and prediction
  - Intermediate layers latently encode phonetic & phonological information
- Can then fine-tuned for a specific task (speech recognition, emotion classification)
- However has *rarely* (once!) been tested as a tool for **phonetic annotation**



Dieck et al. (2022), de Heer Kloots & Zuidema (2024), Kim & Gendrot (2024)

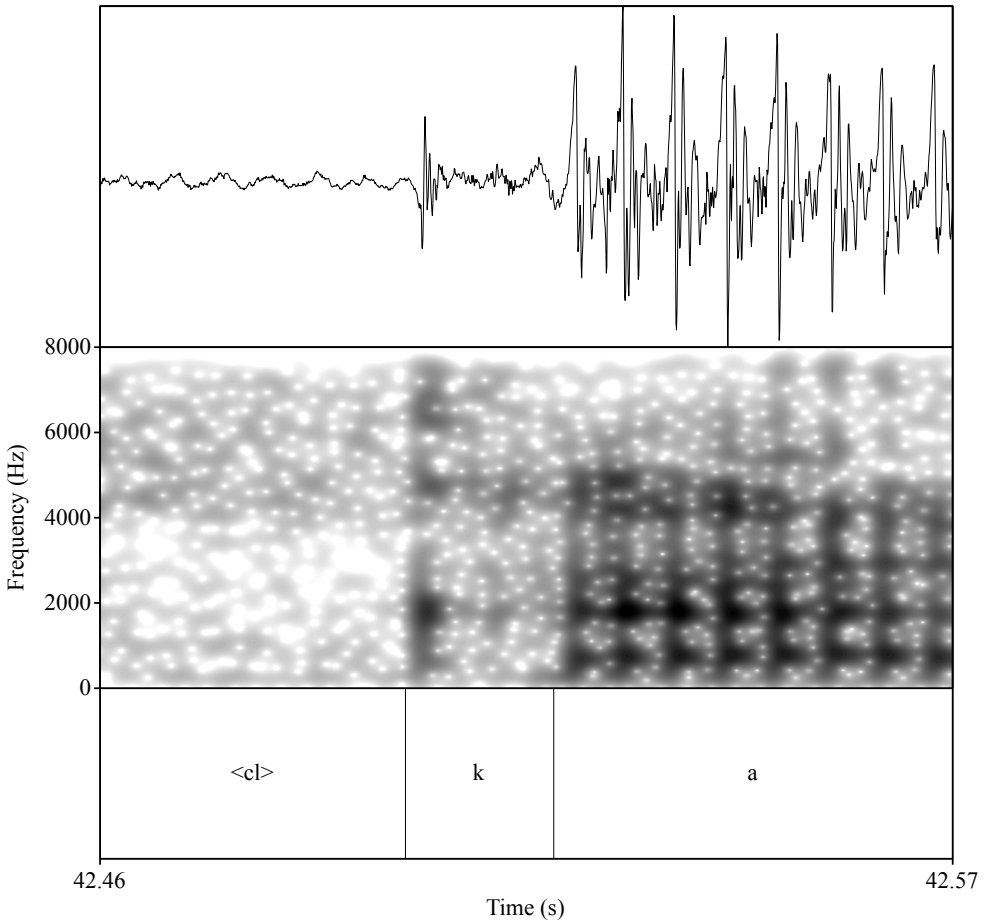
Research Questions

- RQ1: *How well can a pre-trained neural network model (wav2vec 2) predict stop burst presence/absence in:*
  - *Clean, homogeneous, manually-corrected data (experiment 1);*
  - *Noisy, heterogenous, largely-uncorrected data (experiment 2)?*
- RQ2: *How much data is needed for good predictive accuracy?*
- RQ3: *Do predicted stop realisation patterns compare with observed data?*

Experiment 1: Data

大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics
NINJAL

- *Corpus of Spontaneous Japanese (CSJ)*
  - 45 hours (500k words), 137 speakers
  - Spontaneous/semi-spontaneous monologues (+ some conversation)
  - Extensive manual correction
    - segmental boundaries & labels)
    - **stop burst presence/absence**
  - 180k stops
- Take 40k subset of data (20k voiced/ voiceless)
- 80% (32k stops) used for training, 20% (8k) used for evaluation

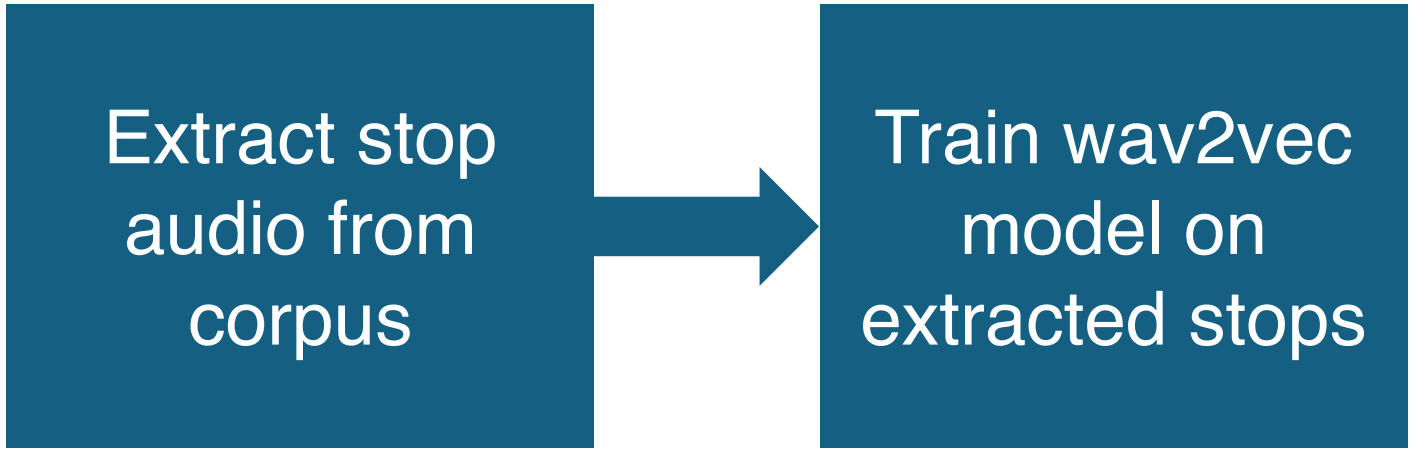


Maekawa et al. (2000), Koiso et al. (2014), Maekawa (2018)

Experiment 1: Training

- Extract audio of each stop, based on start/end times within CSV
  - Include 10ms 'context' on either side
  - Resample to 16kHz
- Write new CSV with wav filename + burst presence/absence label
- Take 40k stop subset (20k voiceless/voiced)

Experiment 1: Training



- Train wav2vec 2.0 to classify stop burst presence/absence using HuggingFace Transformers library in PyTorch
  - 4 different wav2vec2 base models (base-960h, large-960h, lv60, xslr)
- Train for 10 epochs/rounds with 40GB NVIDIA A100 GPU
- Training/evaluation metrics logged with comet.ml

Paszke et al. (2019), Wolf et al. (2020)

NAME	TAGS	SERVER END TIME	EXPERIMENT KEY	DURATION	
CSJ-base-40000		11/18/24 10:07 PM	e9f9cfa96836...	00:08:34	Register model | Reproduce

- Panels
- Code
- Hyperparameters
- Metrics
- Graph definition
- Output
- System metrics
- Installed packages
- Notes
- Graphics
- Audio
- Text
- Confusion matrix
- Histograms
- Other

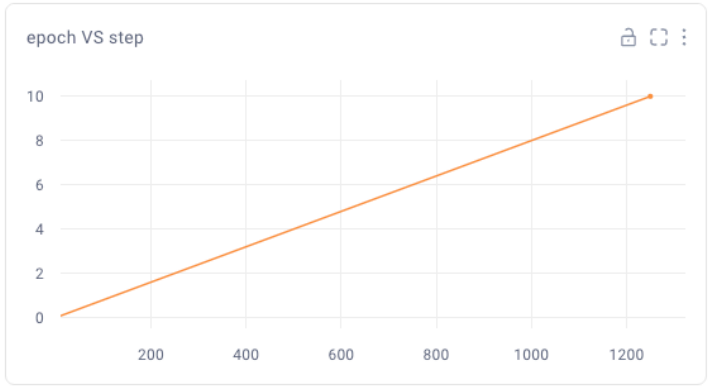
Search by name: Search panel (regex)

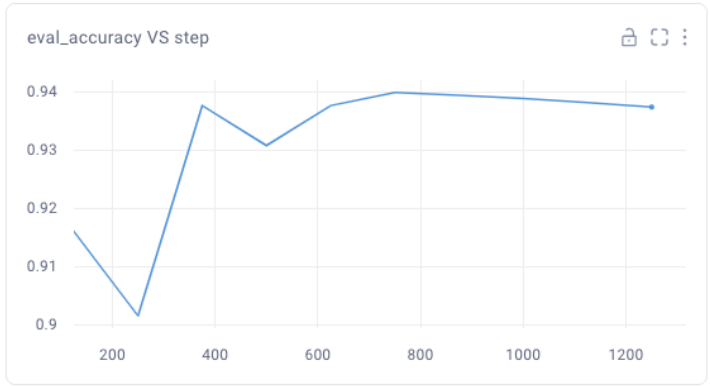
X-axis: Step

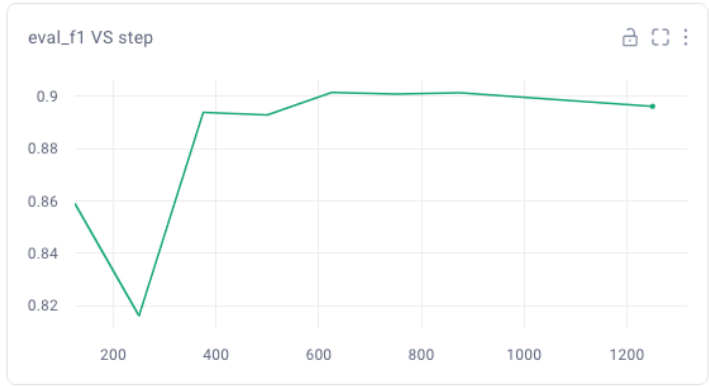
Y-axis smoothing: 0

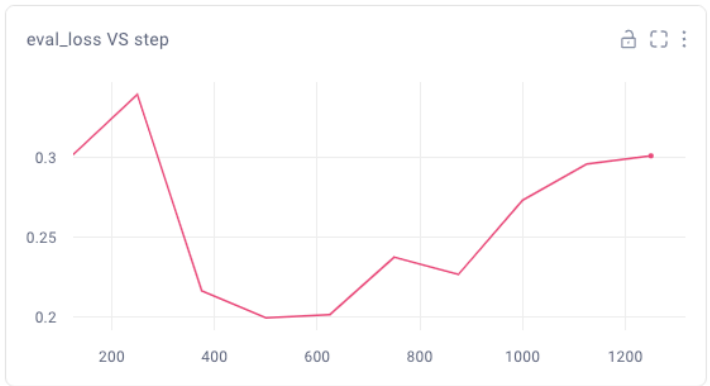
Metrics (15)

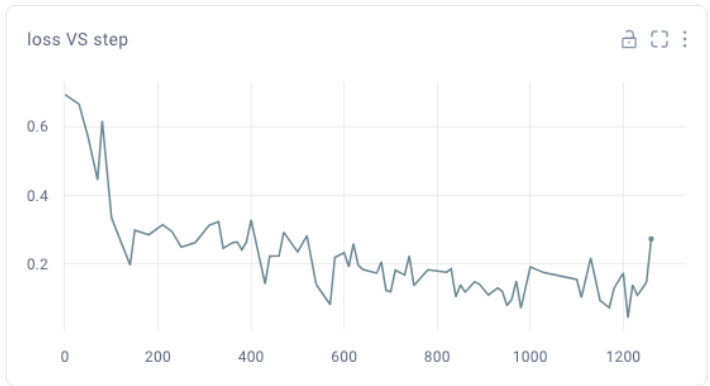
Add panel

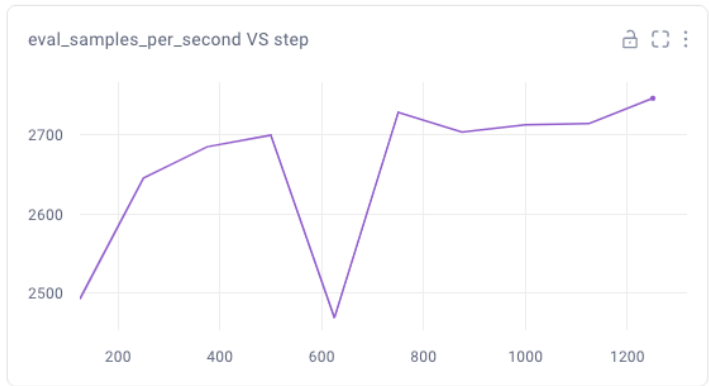












Experiment 1: Results

- All 4 models achieve 91-94% prediction accuracy
- *What about noisier, less-corrected data?*

Experiment 2: Data

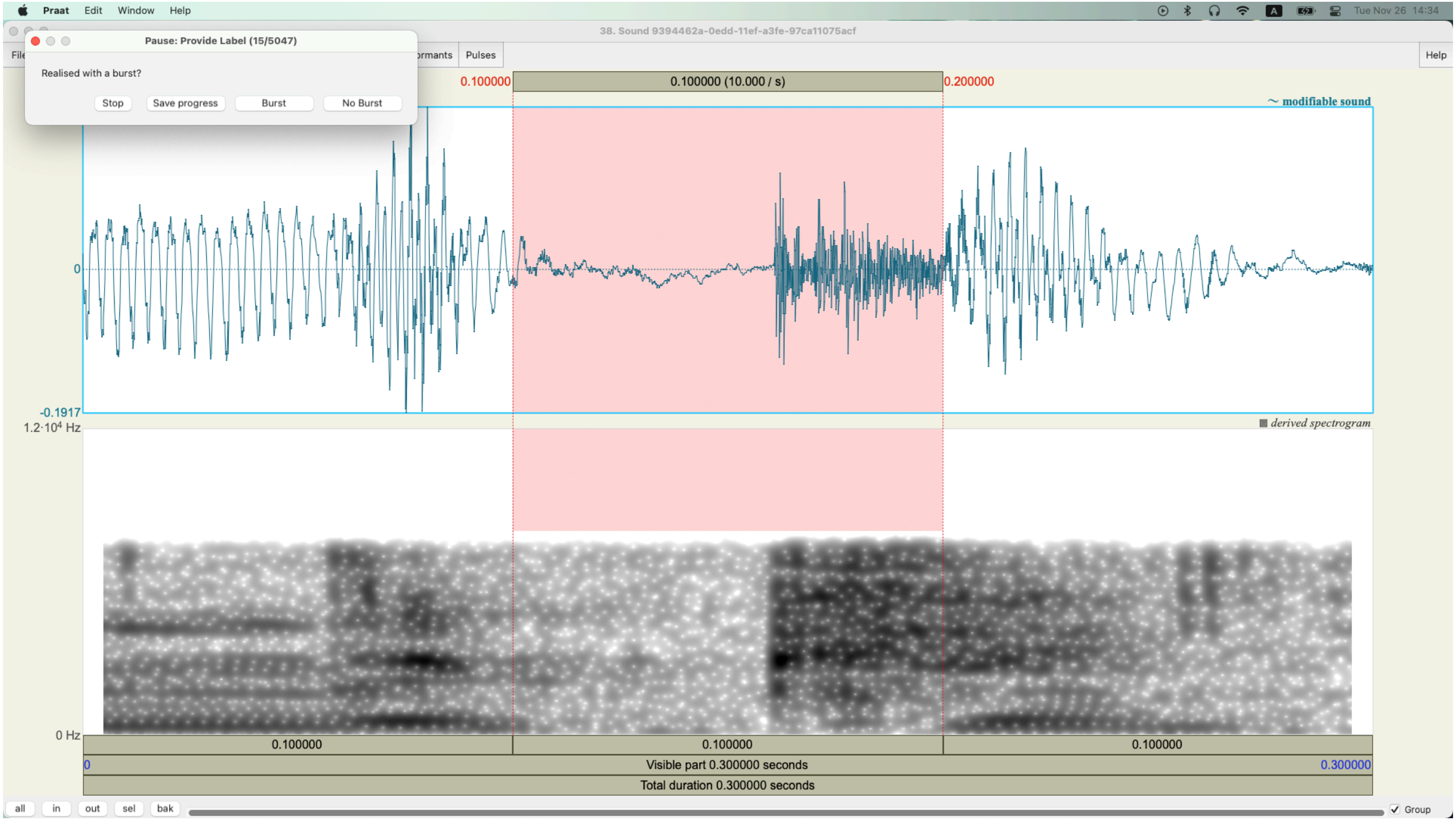
SPADE
SPeech Across Dialects of English

- SPeech Across Dialects of English (SPADE) project:
  - ~40 public & private English speech corpora
  - ~30 different British & North American English dialects
  - Range of speech styles, recording time, and context
  - Force-aligned with minimal manual correction

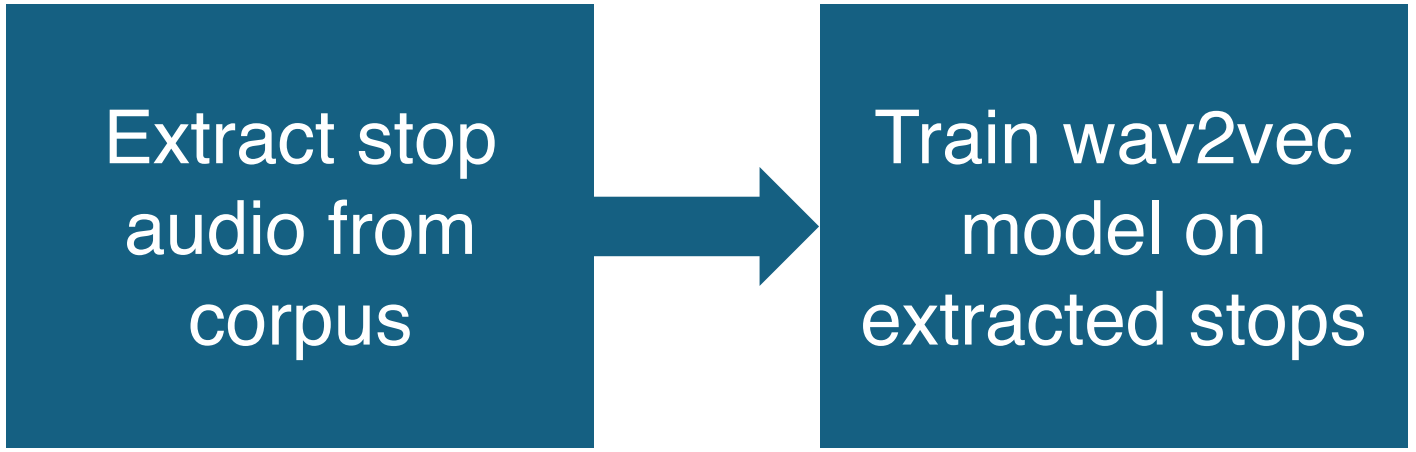
https://spade.glasgow.ac.uk/the-spade-consortium/

Experiment 2: Annotation

- Extract word-initial intervocalic stop audio from all SPADE corpora
- Include 100ms 'context' on either side, to help with manual annotation
- Resample to 16kHz
- 55k stop burst presence/absence annotated by JT using custom Praat script
    - 5-10% per corpus * voicing, up to 1000 stops

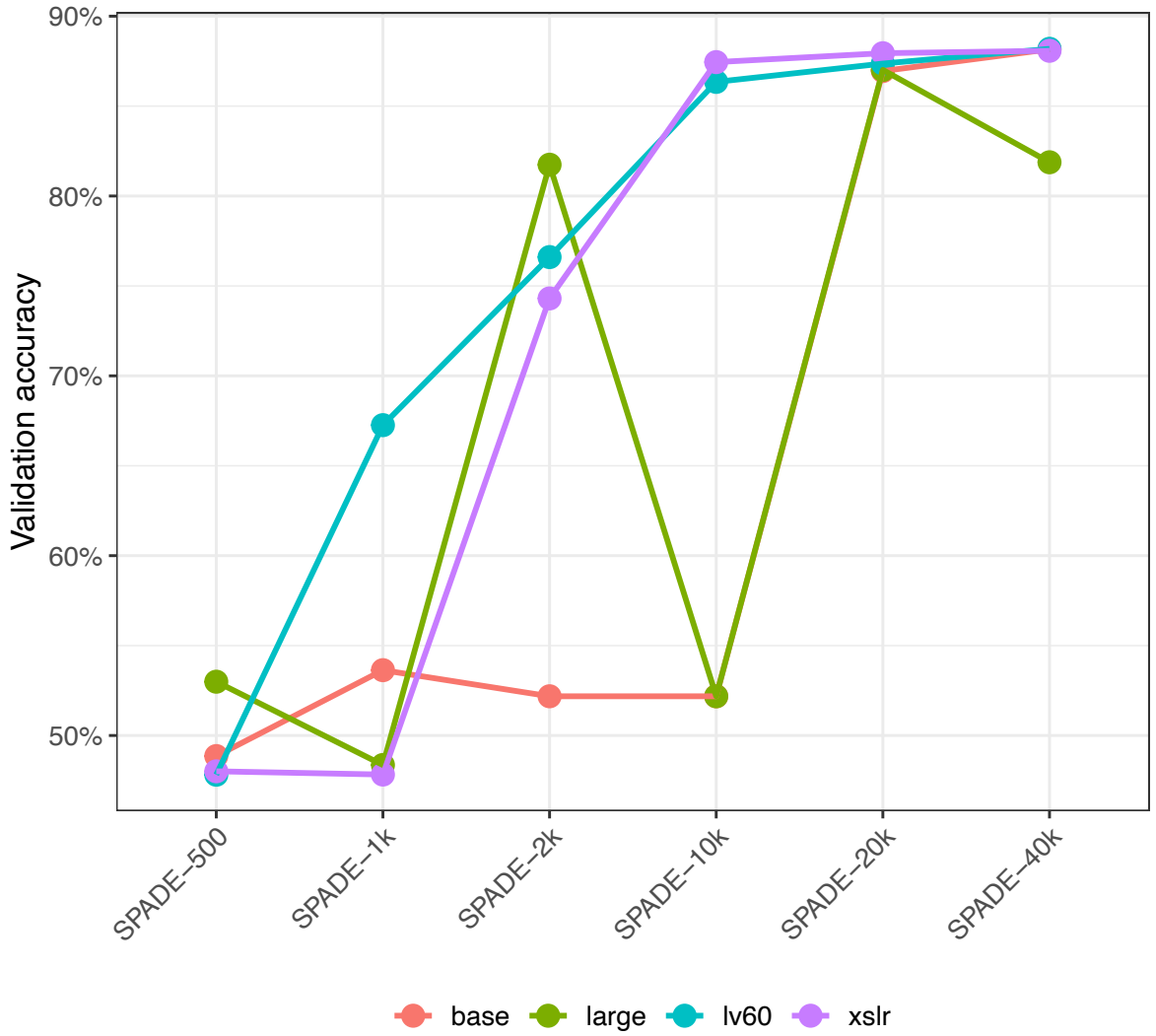

Praat Edit Window Help
Tue Nov 26 14:34
38. Sound 9394462a-0edd-11ef-a3fe-97ca11075acf
Pause: Provide Label (15/5047)
Realised with a burst?
Stop
Save progress
Burst
No Burst
Pulses
Help
0.100000
0.100000 (10.000 / s)
0.200000
~ modifiable sound
0
-0.1917
1.2·10⁴ Hz
derived spectrogram
0 Hz
0.100000
0.100000
0.100000
0
Visible part 0.300000 seconds
0.300000
Total duration 0.300000 seconds
all
in
out
sel
bak
Group

Experiment 2: Training



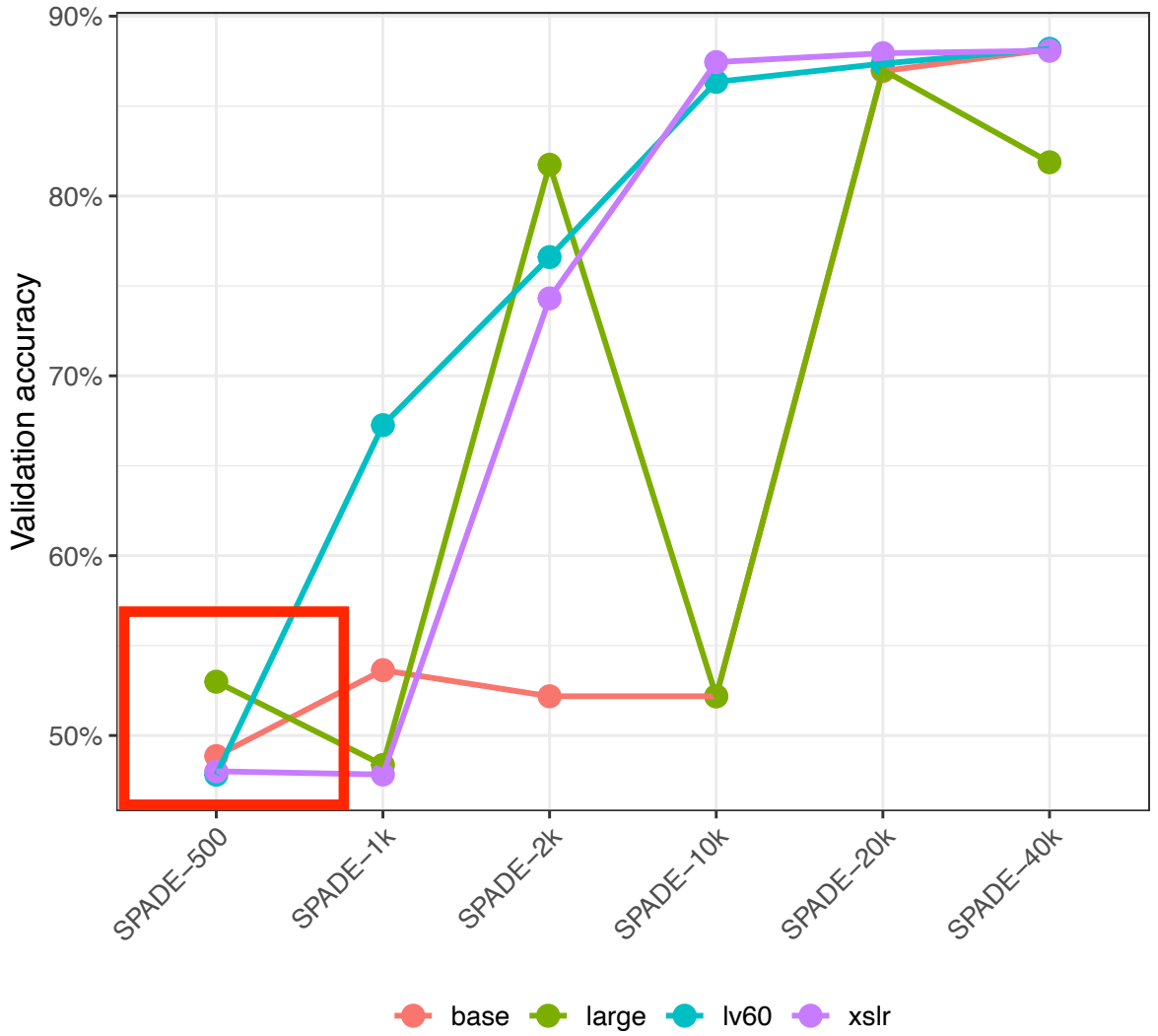
- Re-extract 55k stops with 10ms window
- Take 11k annotated stops as validation set
- *How much data is needed for good results (RQ2)?*
- Train models with different amounts of SPADE data:
  - 500, 1000, 2000, 10k, 20k, 40k

Experiment 2: Results



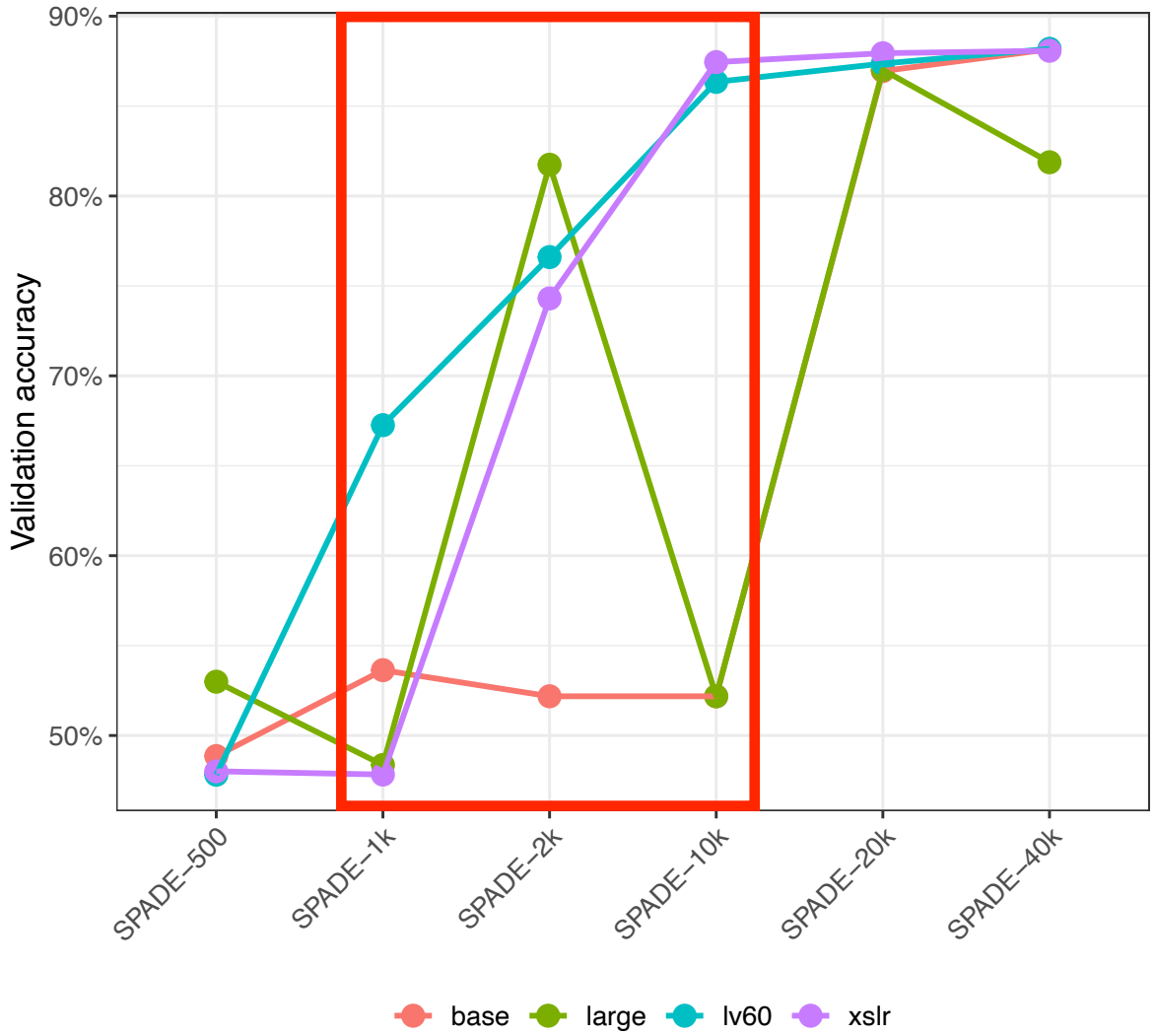
Experiment 2: Results

- Models fail to learn with 500 annotated stops



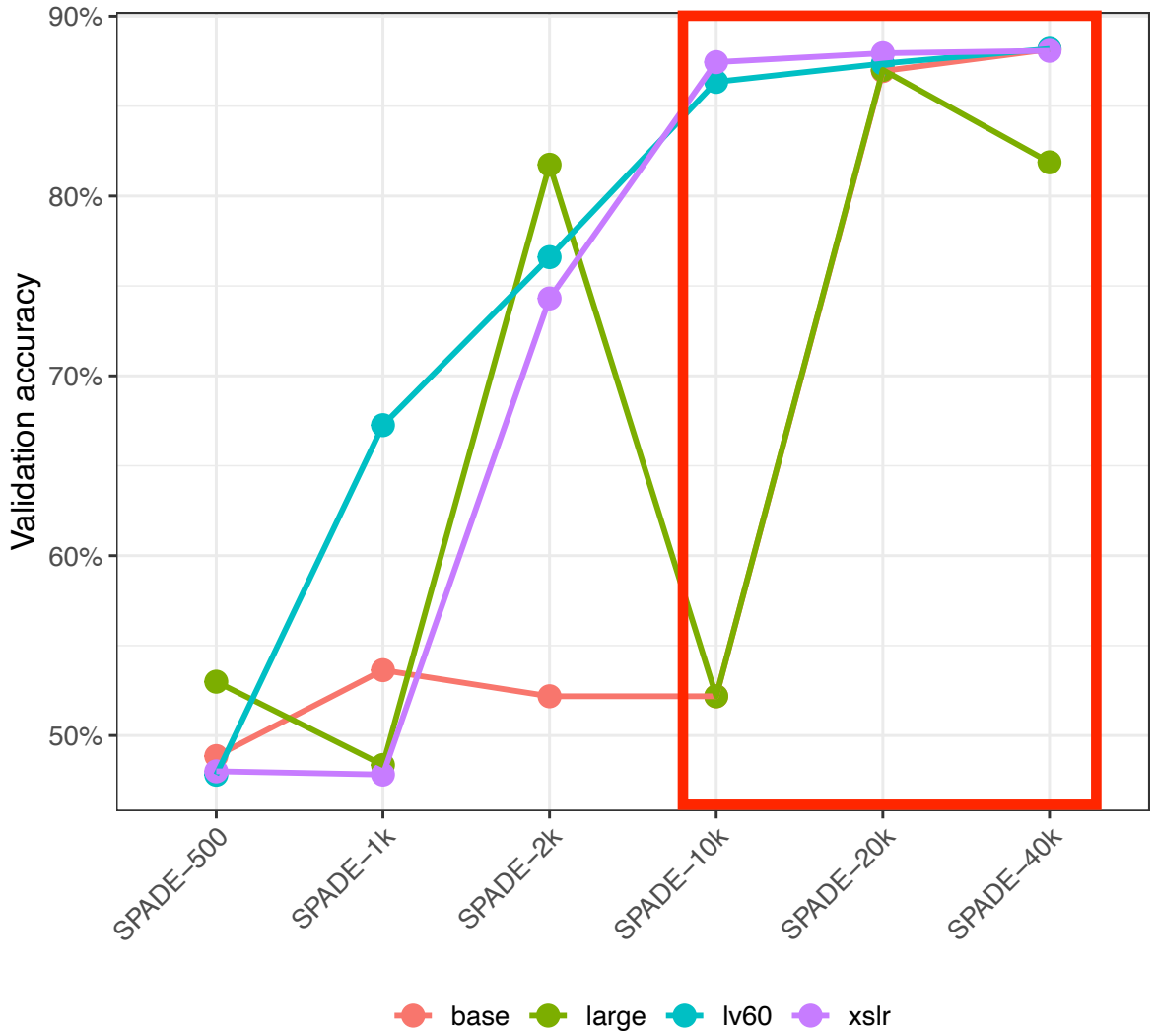
Experiment 2: Results

- Models fail to learn with 500 annotated stops
- Performance at 1k-10k dependent on specific base model



Experiment 2: Results

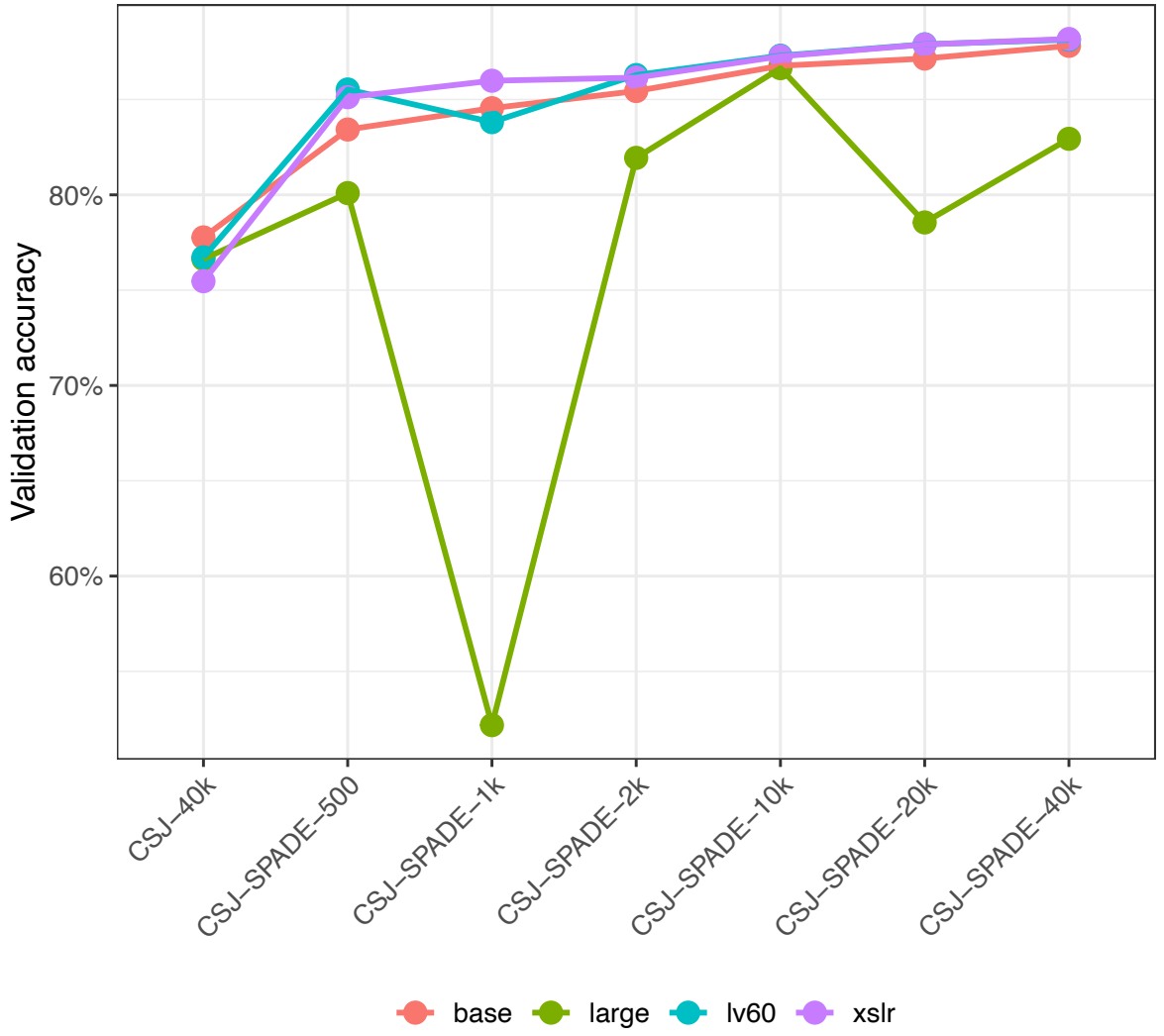
- Models fail to learn with 500 annotated stops
- Performance at 1k-10k dependent on specific base model
- Accuracy plateaus at ~88% for models trained on 10k+ stops



Experiment 2.5: CSJ -> SPADE

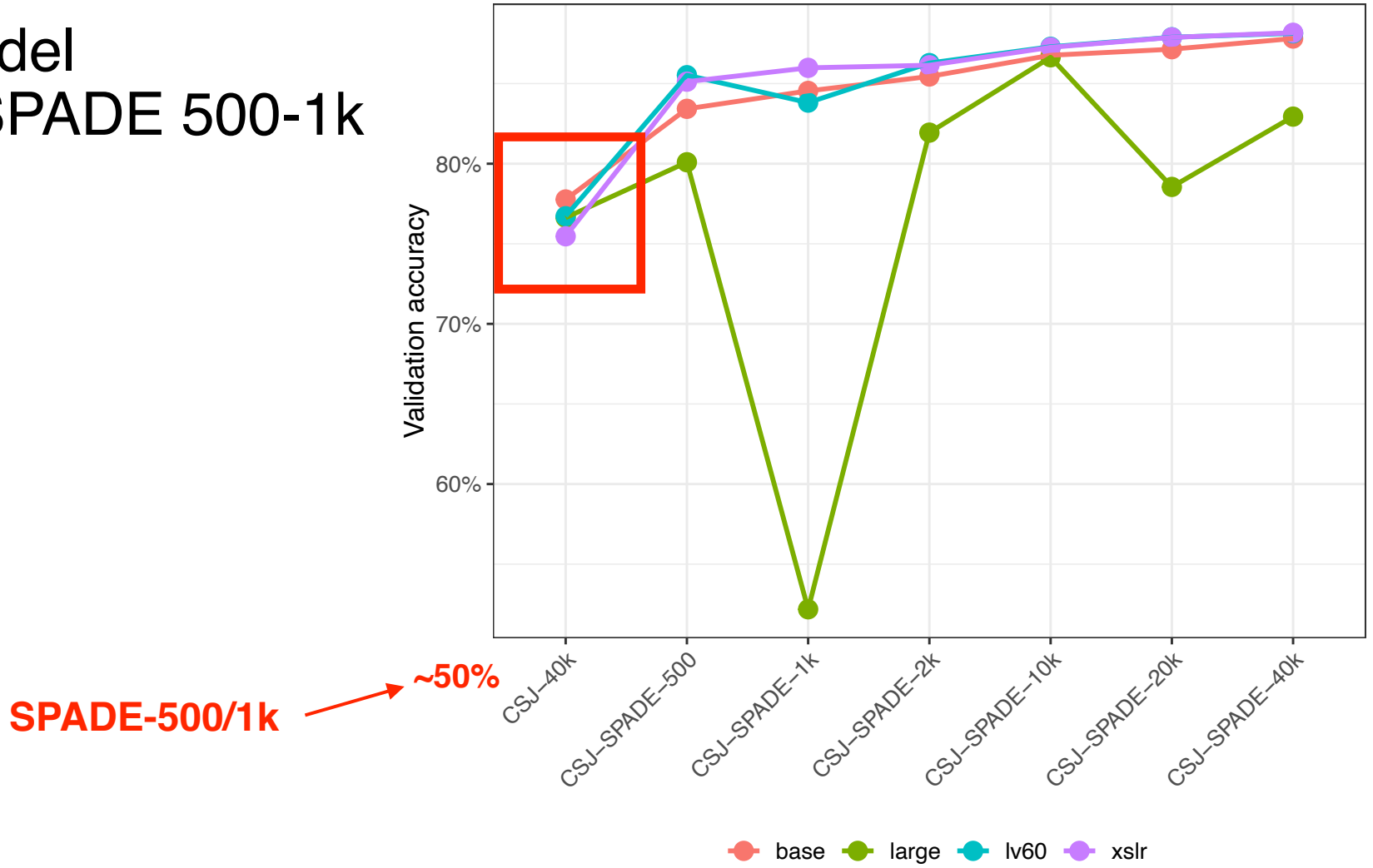
- *Is it possible to attain similar accuracy with less data by bootstrapping from another fine-tuned model?*
- Train same {500, …, 40k} SPADE subsets using CSJ models (Experiment 1) as the base

Experiment 2.5: Results



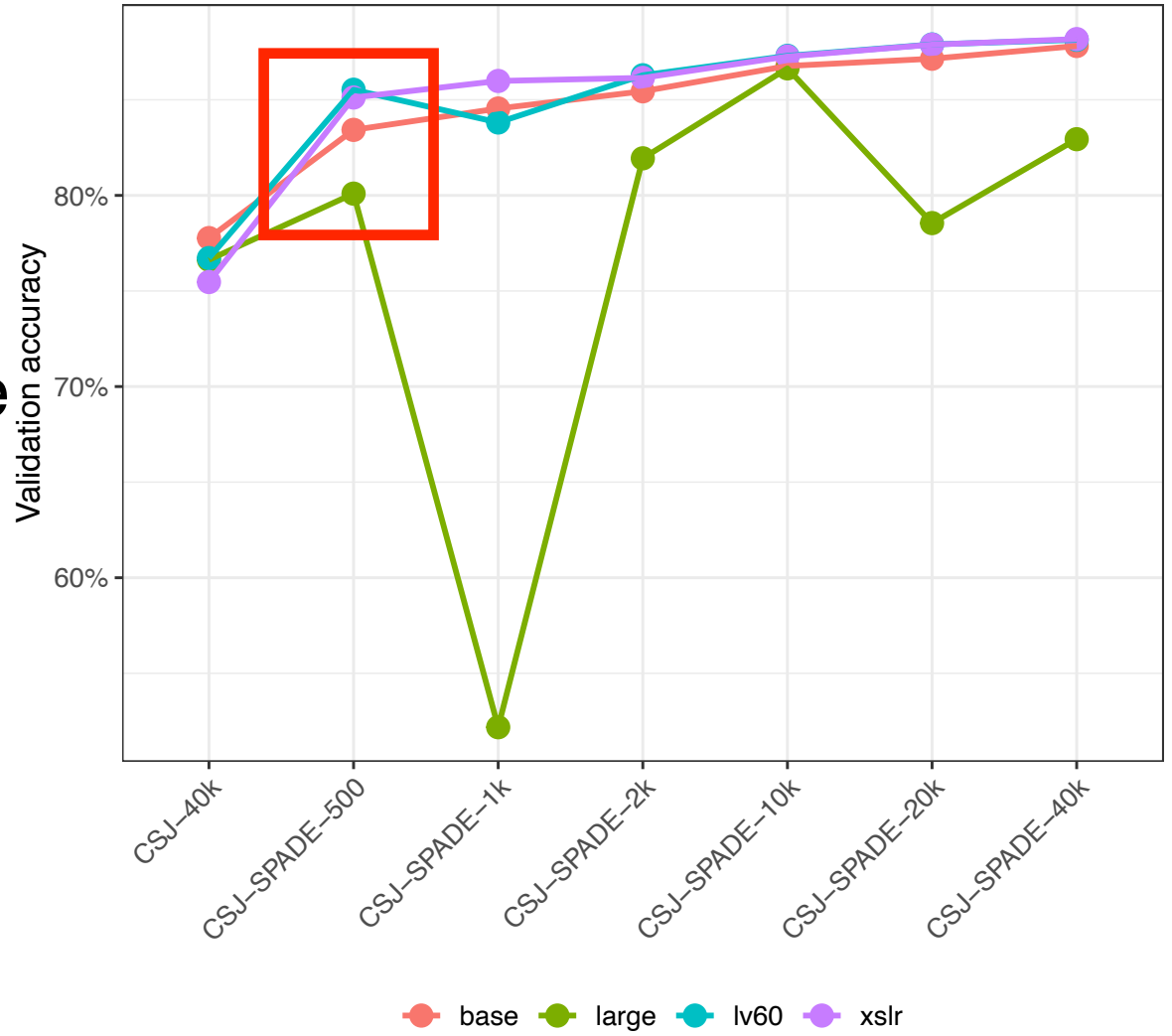
Experiment 2.5: Results

- Base CSJ model outperforms SPADE 500-1k models



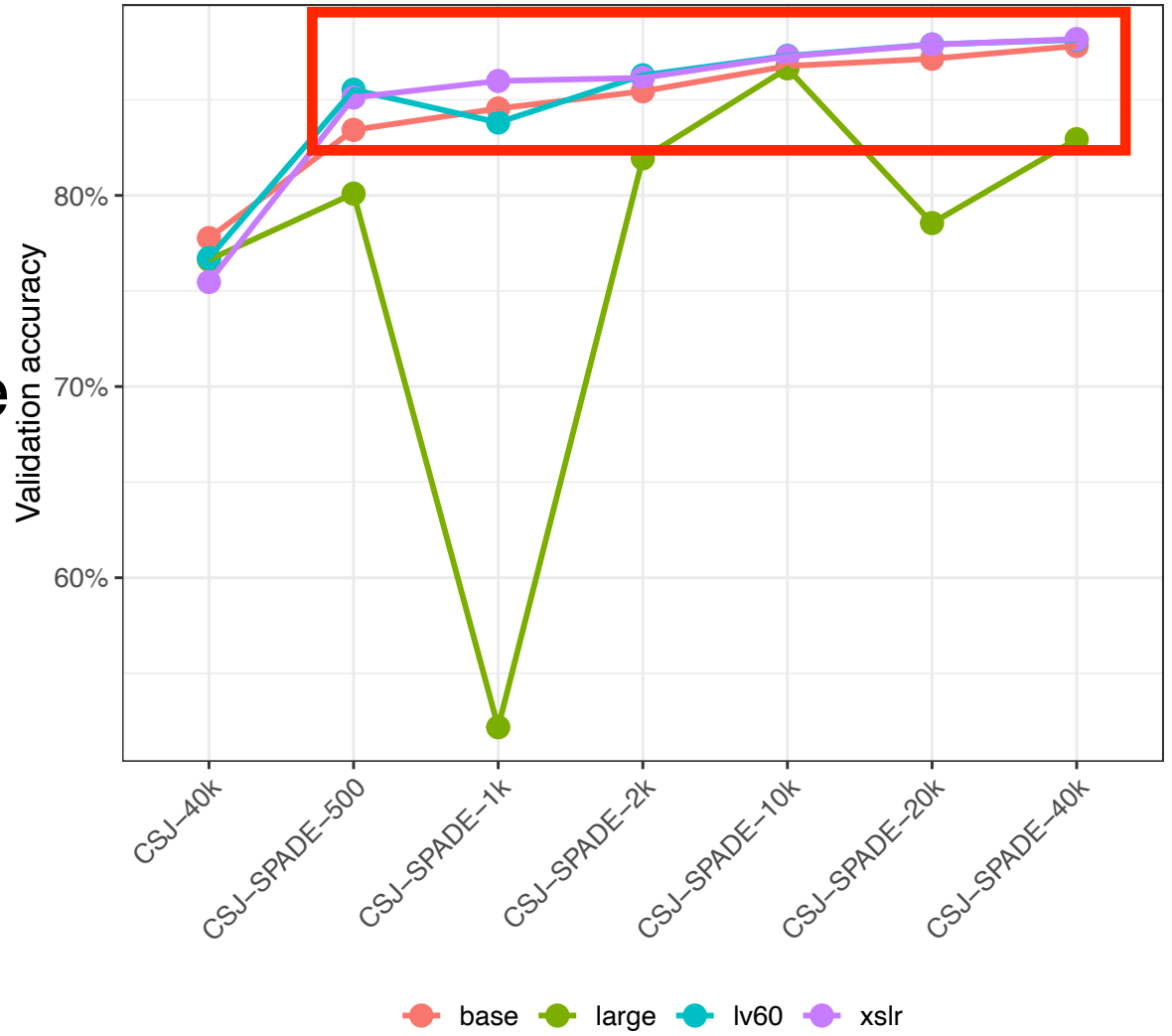
Experiment 2.5: Results

- Base CSJ model outperforms SPADE 500-1k models
- Finetuning with 500 SPADE stops increases performance by ~10%
  - 500 model accuracy within ~5% of 40k model



Experiment 2.5: Results

- Base CSJ model outperforms SPADE 500-1k models
- Finetuning with 500 SPADE stops increases performance by ~10%
  - 500 model accuracy within ~5% of 40k model
- Less variability between models (except for large-960h)



RQ3: Analysis of predicted burst presence/ absence

- *How similar are predicted patterns of stop reduction to those from hand annotations?*
- Compare probability of burst presence by **phonological voicing** and **phone duration** between:
  - Manual annotations
  - lv60 500-stop model predictions
  - xslr 40k-stop model predictions

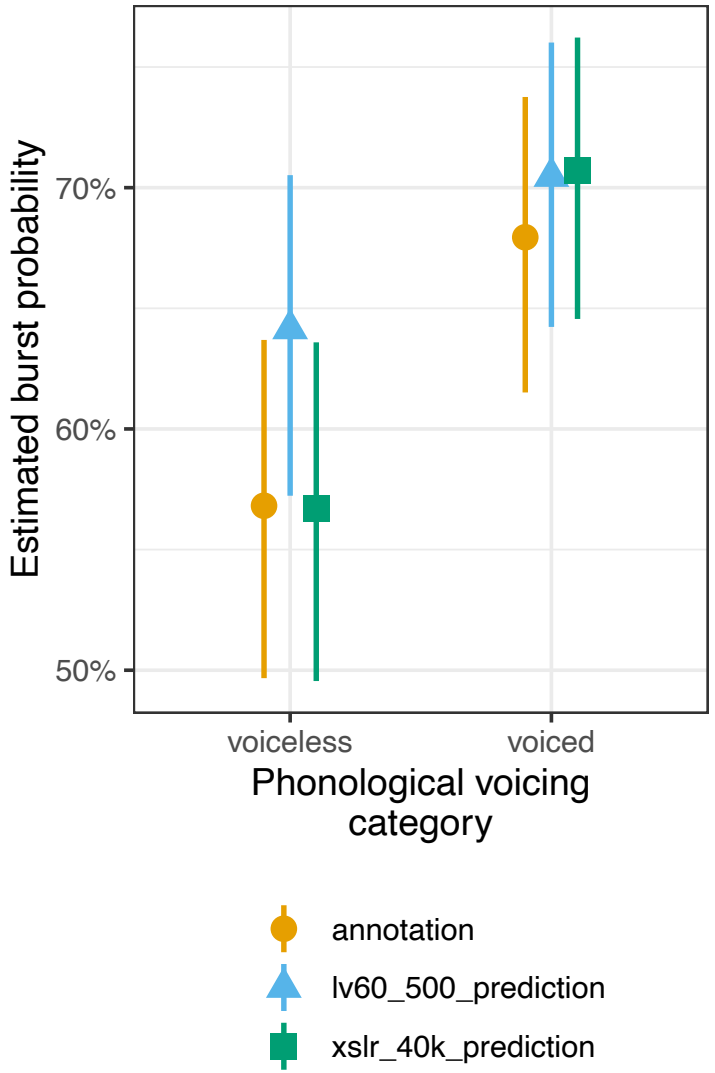
RQ3: Analysis of predicted burst presence/ absence

- Fit logistic GAMM of stop burst likelihood
  - (Log) phone duration smooth term
  - Parametric + smooth term by annotation type (manual vs lv60-500-prediction vs xlsr-40k-prediction)
  - Parametric + smooth term by voicing (voiced vs voiceless)
  - By-corpus random smooth (n=75)
  - By-speaker random slopes (n=1334)

Wood (2011), Lenth (2023)

RQ3: Analysis of predicted burst presence/ absence

- Small difference between annotations and predictions
  - ~0% difference for voiceless; ~3% for voiced

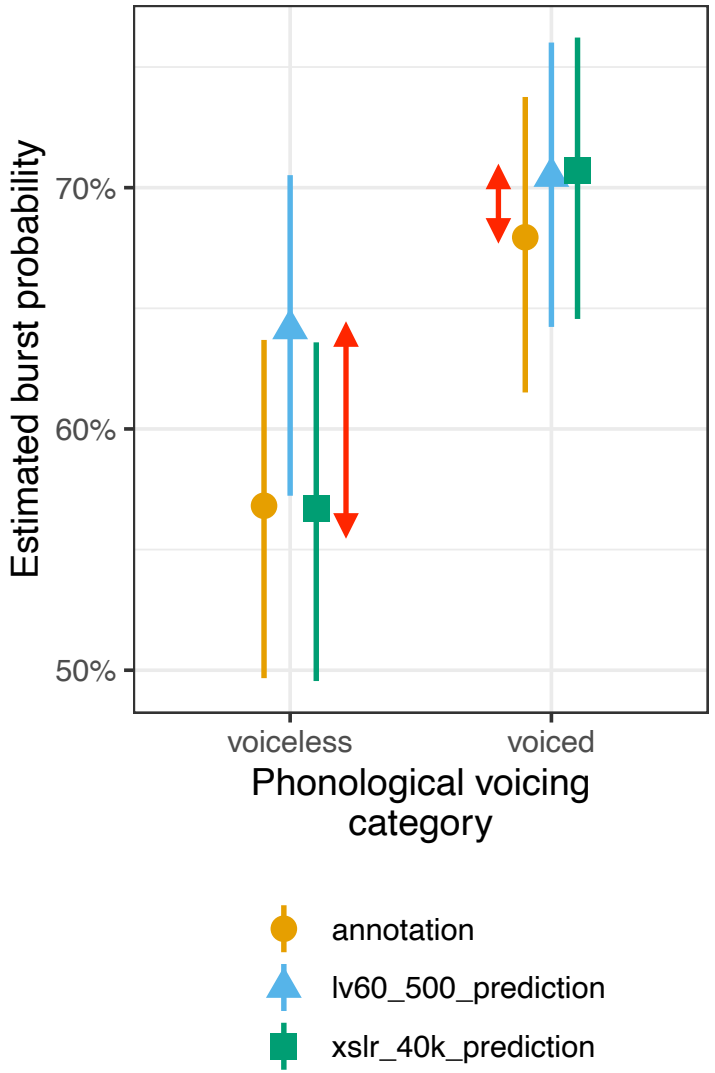


Voiceless: $\Delta_{est} = 0.01$, $t = 0.08$, $p = 0.99$;
Voiced: $\Delta_{est} = -0.13$, $t = -2.76$, $p < 0.05$

RQ3: Analysis of predicted burst presence/ absence

- Small difference between annotations and predictions
  - ~0% difference for voiceless; ~3% for voiced

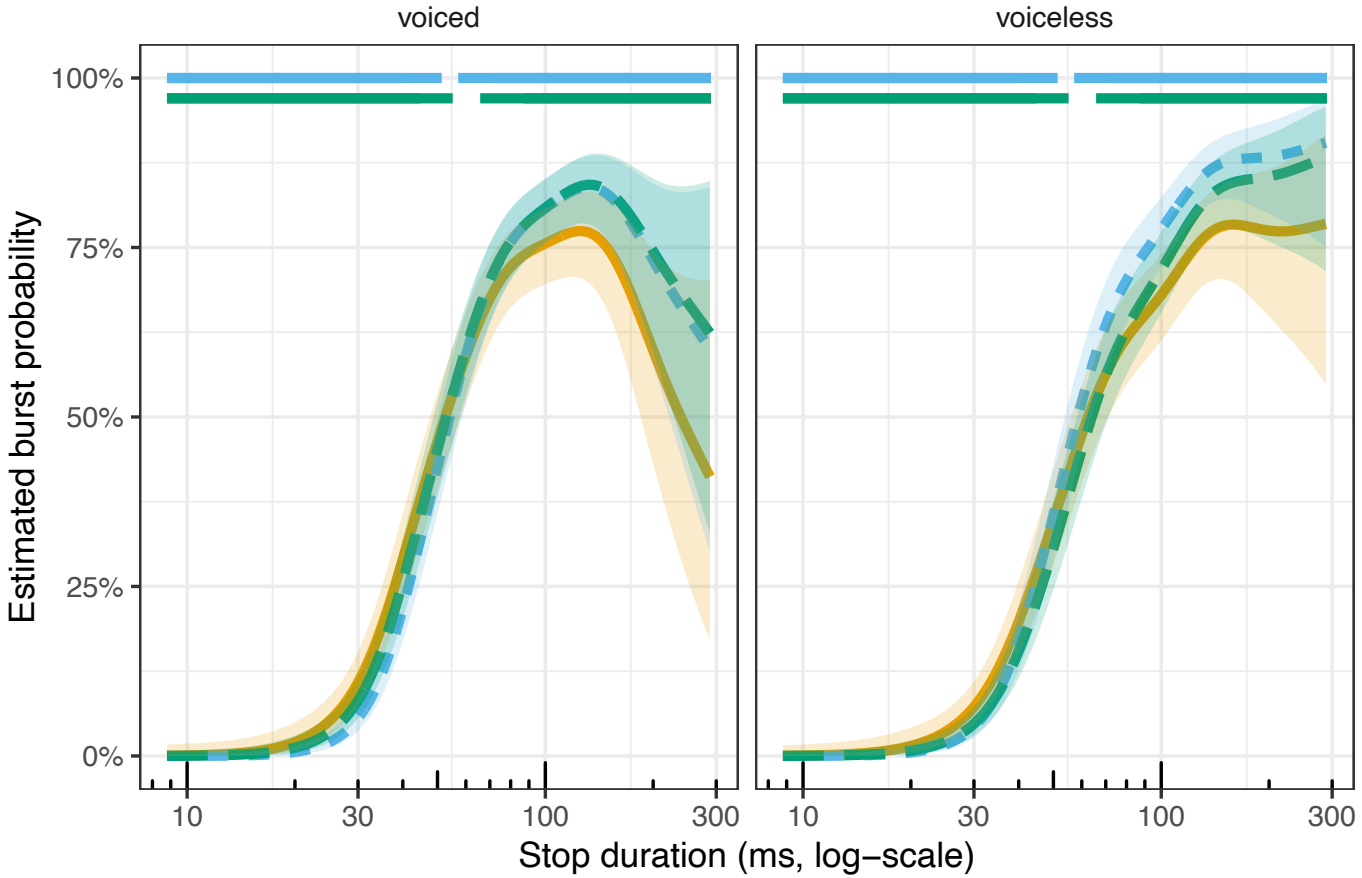
Difference between annotated vs predicted burst probability



Voiceless: $\Delta_{est} = 0.01$, $t = 0.08$, $p = 0.99$;
Voiced: $\Delta_{est} = -0.13$, $t = -2.76$, $p < 0.05$

RQ3: Analysis of predicted burst presence/ absence

- Duration effect on burst probability similar for 10-75ms
- Divergence for very long stop durations
  - Predictions more likely to classify stop with a burst



$\chi^2 (4) = 114.21, p < 0.001$

Summary

- RQ1: *How well can a statistical model (wav2vec2) predict stop burst presence/absence in:*
  - *Clean, homogeneous, manually-corrected data (experiment 1);*
  - *Noisy, heterogenous, largely-uncorrected data (experiment 2)?*

Summary

- RQ1: *How well can a statistical model (wav2vec2) predict stop burst presence/absence in:*
  - *Clean, homogeneous, manually-corrected data (experiment 1);*
  - *Noisy, heterogenous, largely-uncorrected data (experiment 2)?*

- Stop burst prediction accuracy reaches 94% for highly-curated data and 88% for noisy, uncorrected data

Summary

- RQ2: *How much data is needed for good predictive accuracy?*

Summary

- RQ2: *How much data is needed for good predictive accuracy?*
  - Models trained on 500 stops can perform nearly as well as 40k stops, after bootstrapping from a previously-trained model
  - Suggests that information about stop realisation can be transferred *across languages* (e.g. Japanese -> English)

Summary

- RQ3: *Do predicted stop realisation patterns compare with observed data?*

Summary

- RQ3: *Do predicted stop realisation patterns compare with observed data?*
  - Patterns of predicted stop burst largely follow the observed/ annotated patterns
  - Higher burst likelihood for voiceless stops & longer durations
    - Likely due to phonological rules on English voiceless stops (flapping, glottaling)

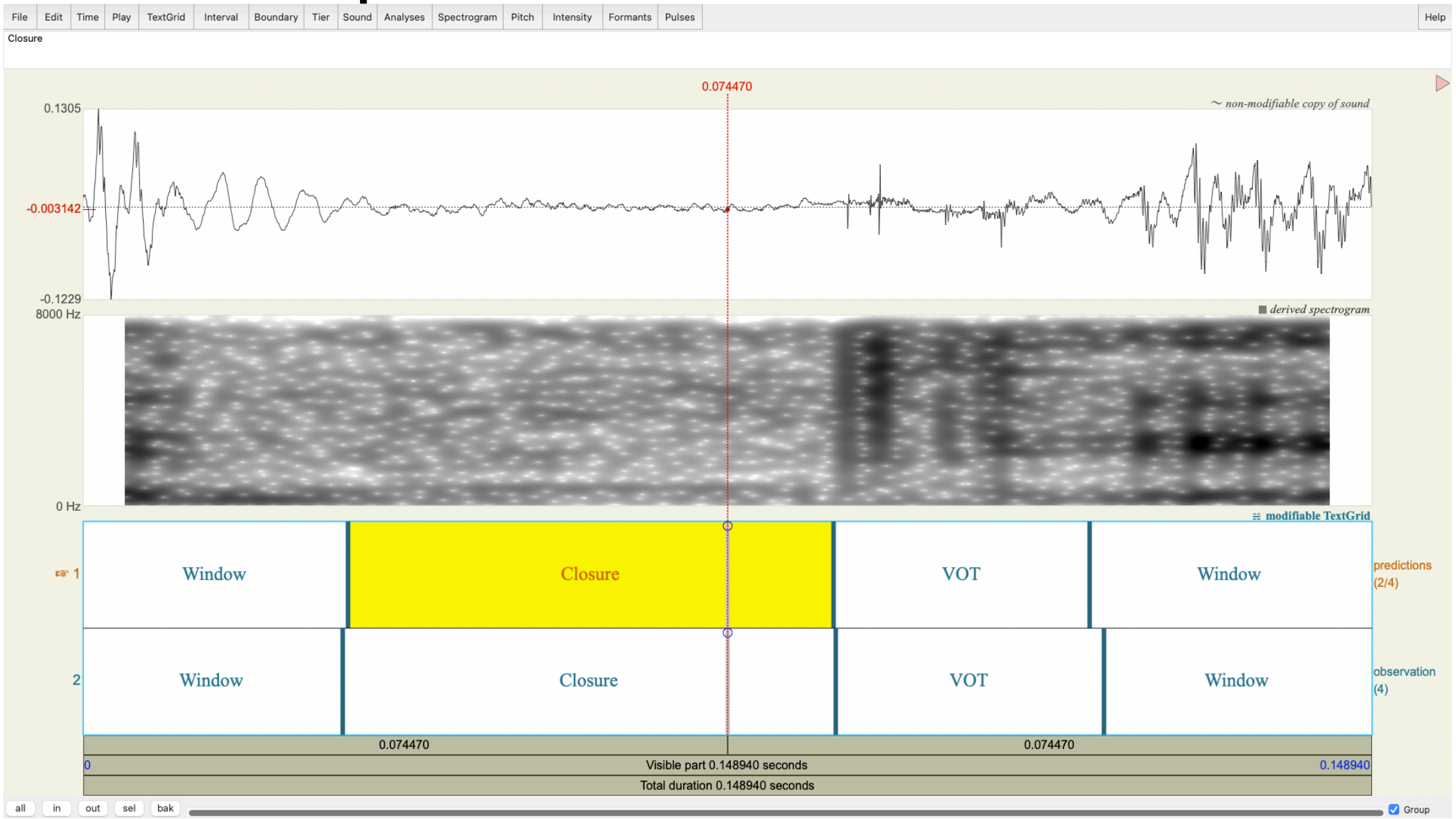
Discussion

- Demonstrates that the realisation of stops can be accurately predicted automatically from both highly
- This can be used as a tool for the (semi-) automatic annotation of stop burst presence/absence, with good performance requiring only a small number of annotated tokens
- Software & models available soon!
  - Core training & prediction (GPU & CPU support)
  - Utility scripts (data extraction & formatting, annotation, combining/ separating, validation, etc)

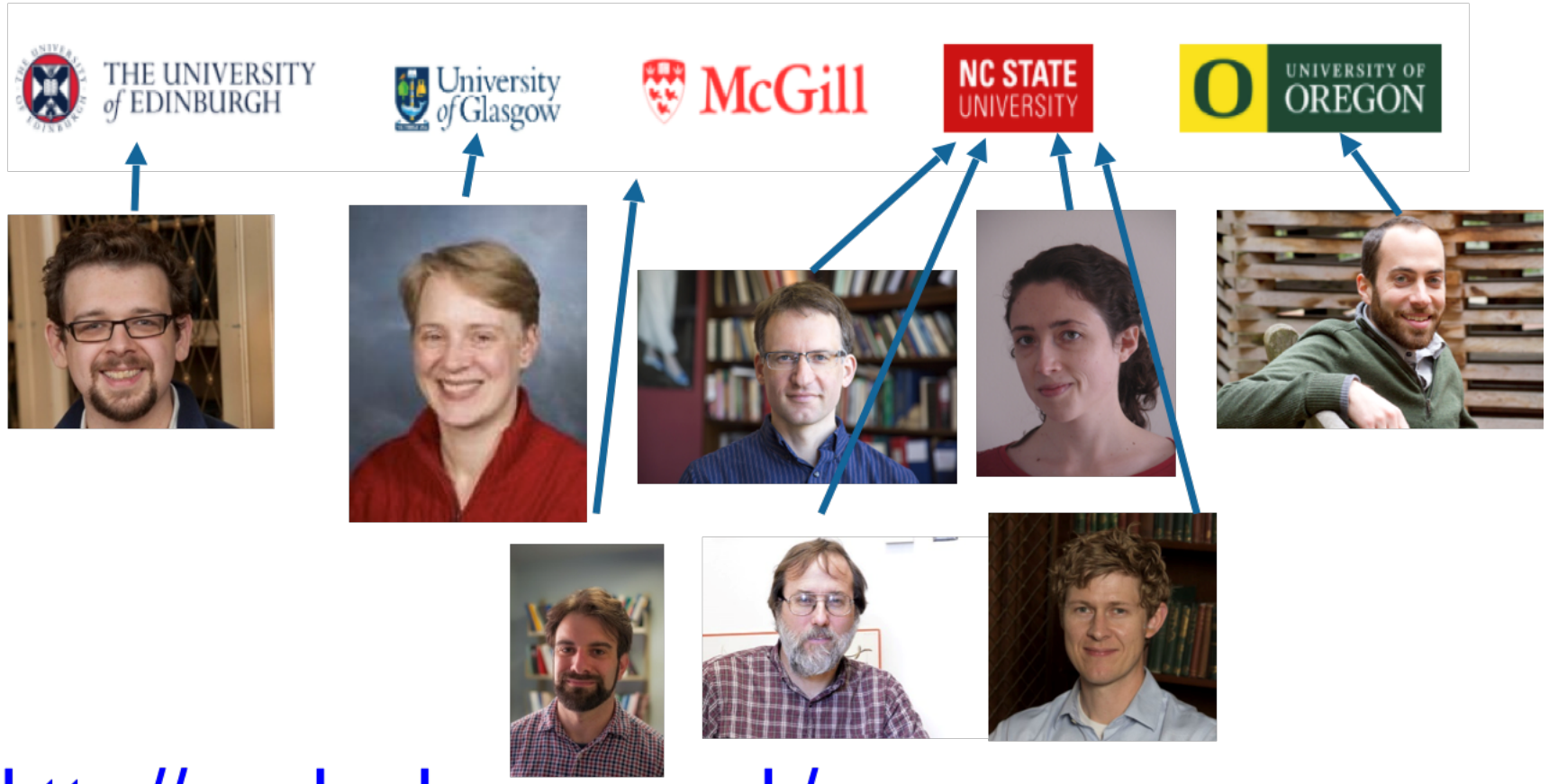
What about temporal annotations?

Is it possible to train wav2vec2 to predict Voice Onset Time (VOT)?

Watch this space…



Thank you!



http://spade.glasgow.ac.uk/

Thank you!


Transformers


PyTorch


comet


The
British
Academy


Trans-Atlantic Platform
Social Sciences and Humanities


DIGGING INTO DATA
CHALLENGE


Calcul Québec


Arts & Humanities
Research Council


E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL


NSERC
CRSNG

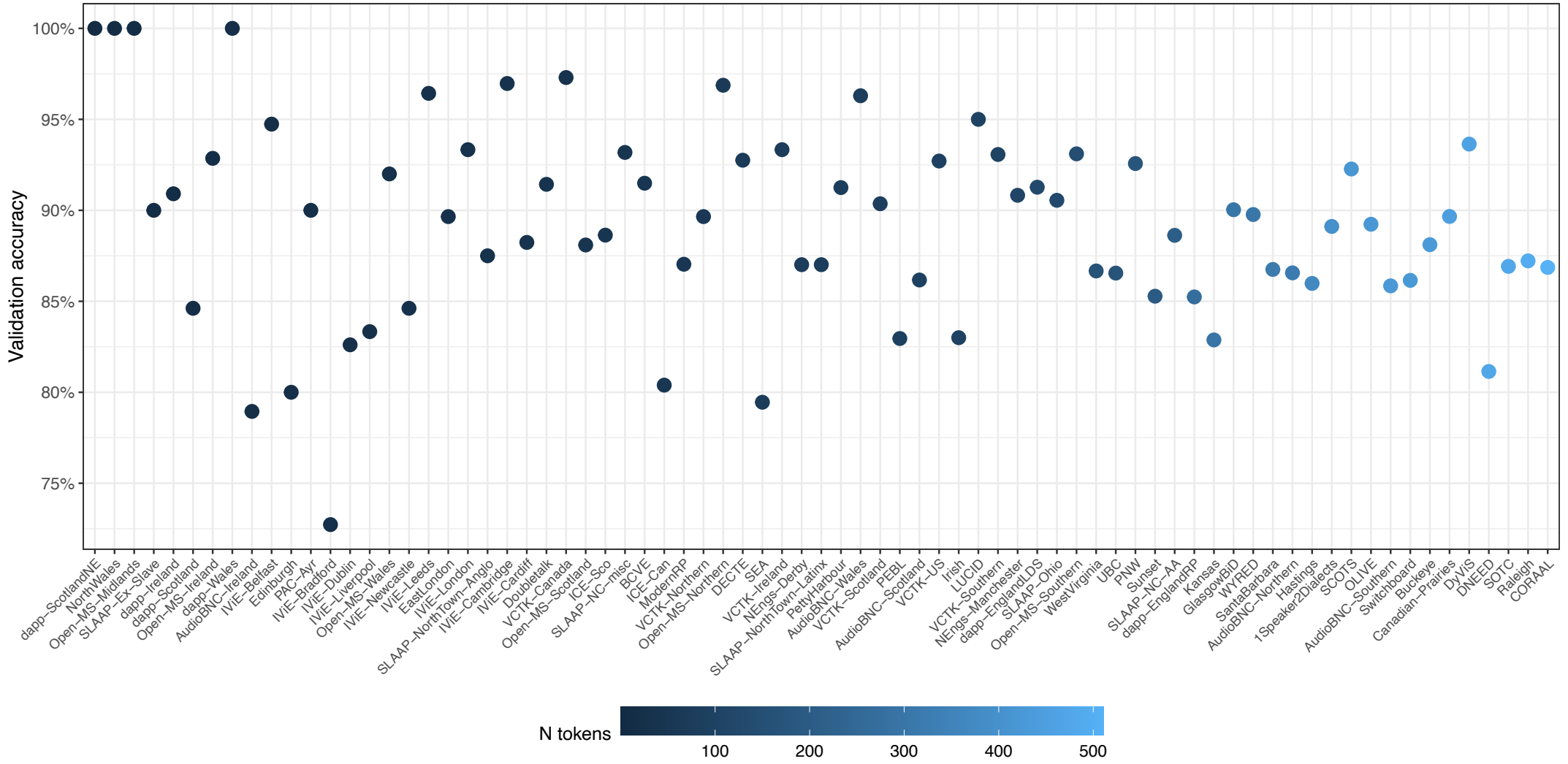

SSHRC CRSH
Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada


NSF

Extra slides: wav2vec base models

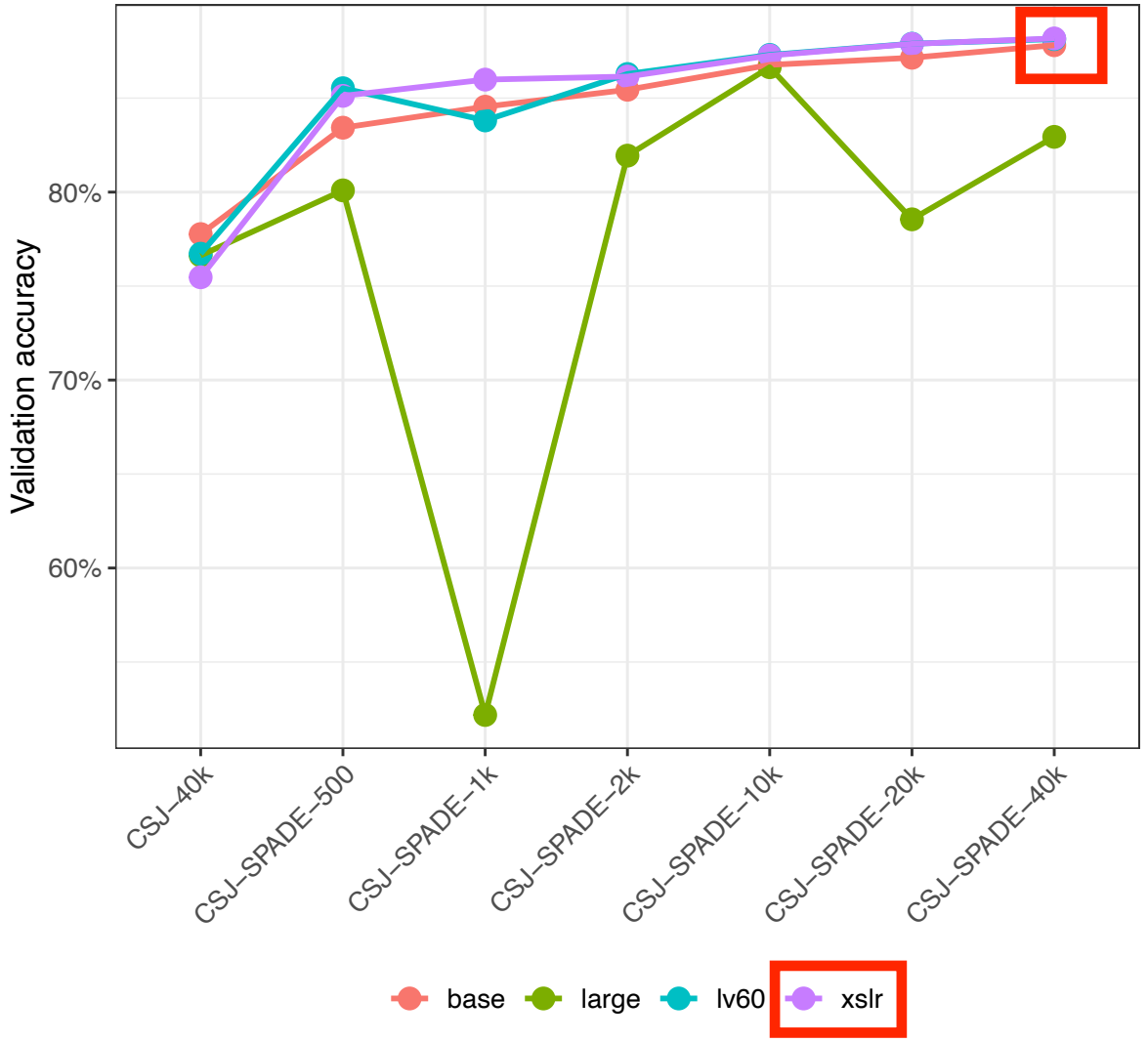
Name	Number of params	Training data	Training Language
Base-960h	35m	Librispeech (960 hours)	English
Large-960h	317m	Librispeech (960 hours)	English
Large-lv60	317m	Librispeech + LibriLight (53 hours)	English
Large-xslr	317m	Multilingual-Librispeech, CommonVoice, BABEL	Multilingual (53 languages)

Experiment 2.5: Results

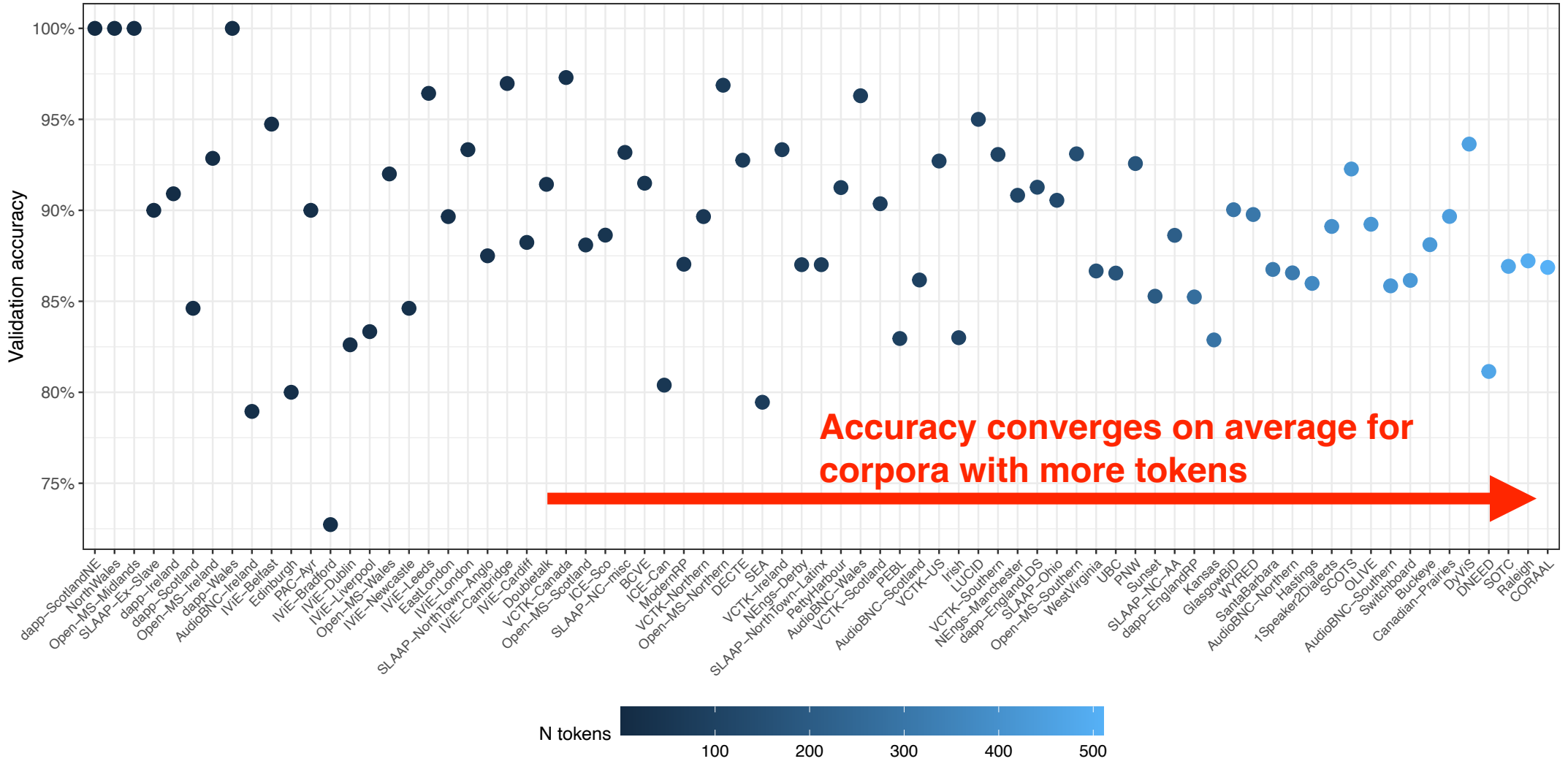


Experiment 2.5: Results

- Use predictions from CSJ-SPADE-40k (xslr) model on 11k SPADE validation set



Experiment 2.5: Results



Extra slides: Comparing 500 & 40k