New Efforts in Large-scale Linguistic Research

Kenneth Church and Jiahong Yuan

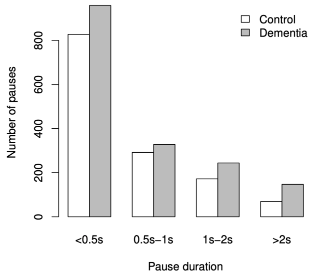
Baidu Research

LDC Workshop for Penn China Research
Nov. 9-10, 2020

Unifying Themes

Found Data	Size (M words)
Audio Books	111.4
SCOTUS	70.0
Audio BNC	7.1
Tedlium	5.7
History	5.0
Presidential	1.5
CommonVoice	0.7

- Forced Alignment of Found Data
  - Input: Audio + Text
  - Output Timestamps: words, phones, silences
- Technologies
  - Machine Learning: Classification/Boosting/ERNIE/BERT
  - Fine-Tuning of language models with pauses (from audio)
    - Audio + Text are better together
- Linguistic Questions
  - Phrase final lengthening:
    - Some ``units’’ are ``longer’’ than ``otherwise’’ in certain ``contexts’’
  - t/d deletion
    - Some ``units’’ are ``deleted’’ in certain ``contexts’’
- Practical Questions
  - Dementia Challenge: Distinguish AD from controls
  - Observation: Pauses are helpful
    - (Intuitively, disfluencies ~ more pauses, laughter, etc.)



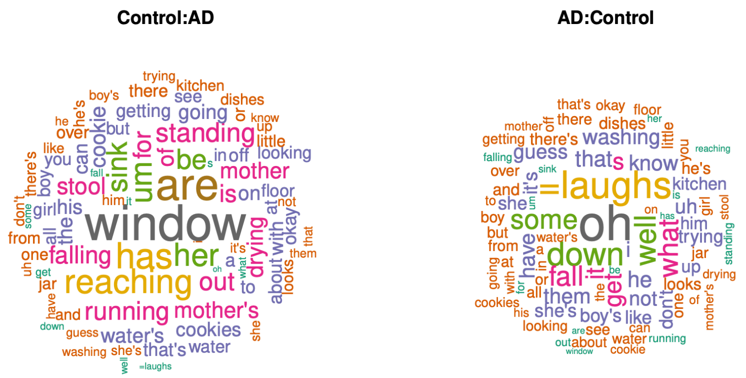
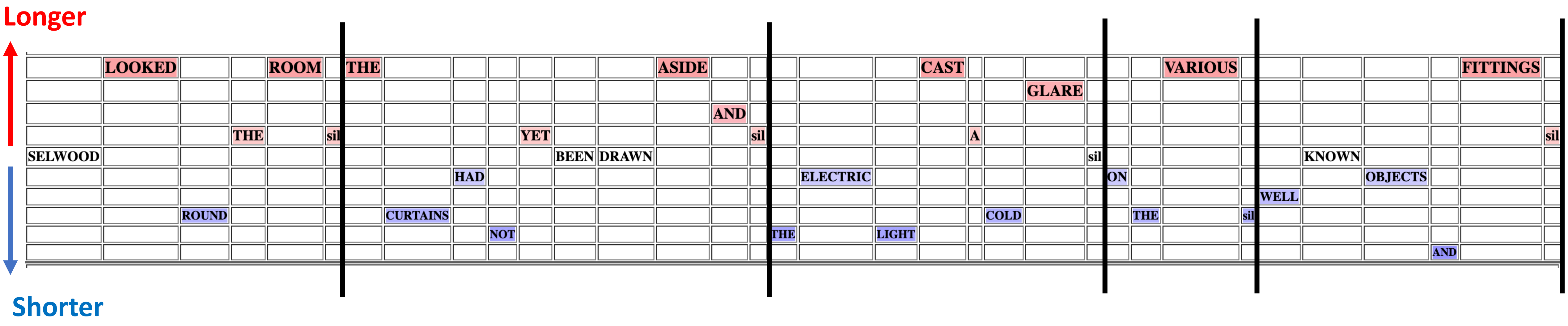


Figure 2: *Subjects with AD have more pauses (in all duration bins).*

Duration Modeling

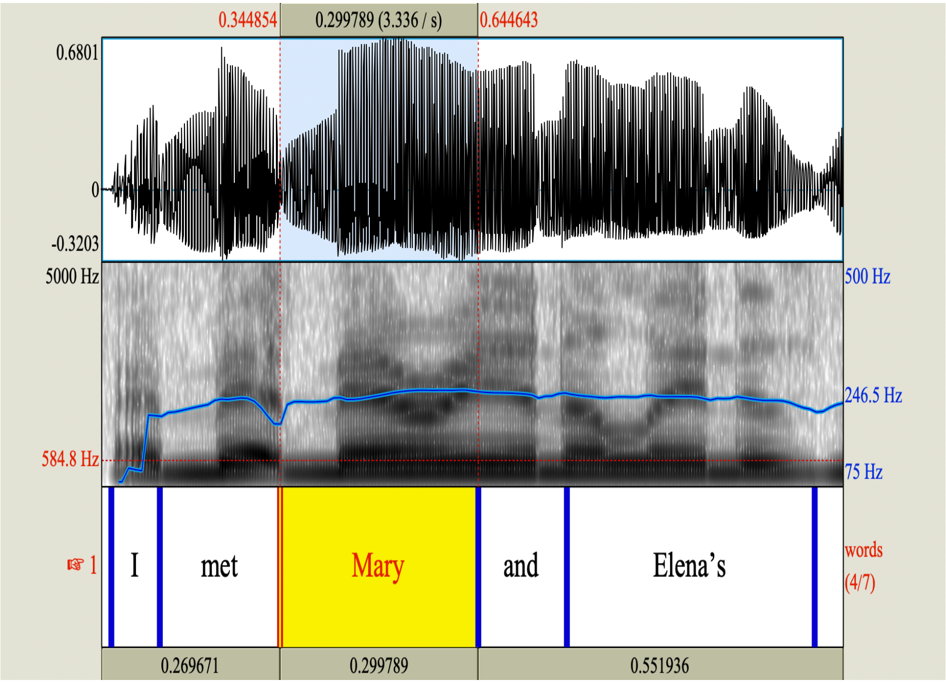


- Phrase final lengthening
  - Words are longer than ``otherwise''
  - before phrase boundary (silence)
- But how do we define ``otherwise''?

0.3 seconds is shorter than ``otherwise''
0.5 seconds is longer than ``otherwise''

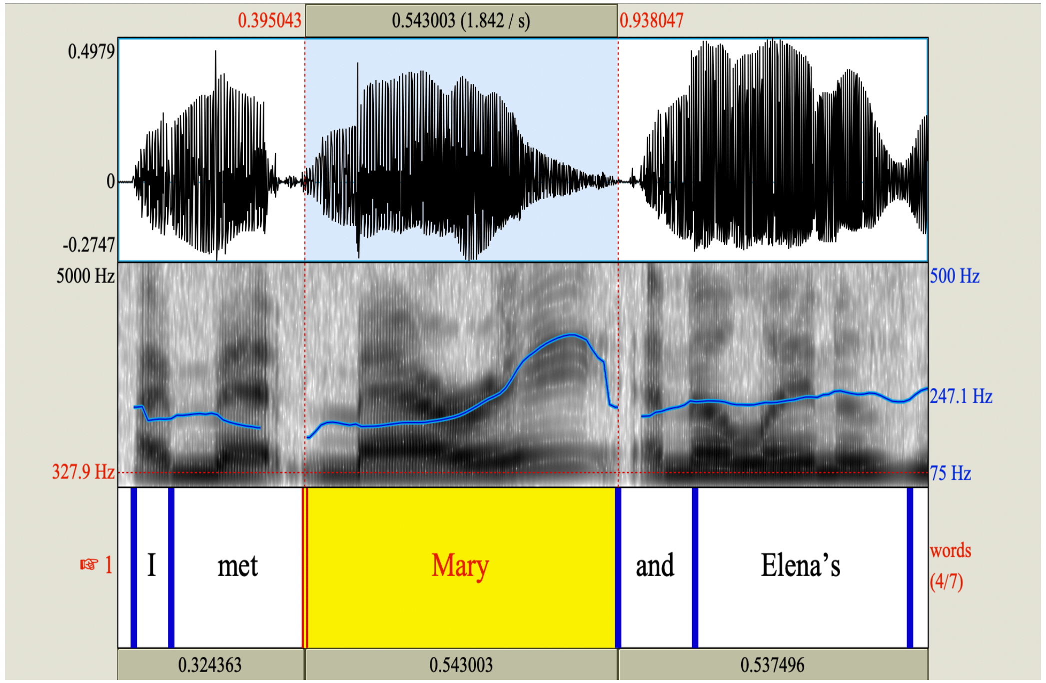
Conjunction: Narrow Scope (0.3 sec)

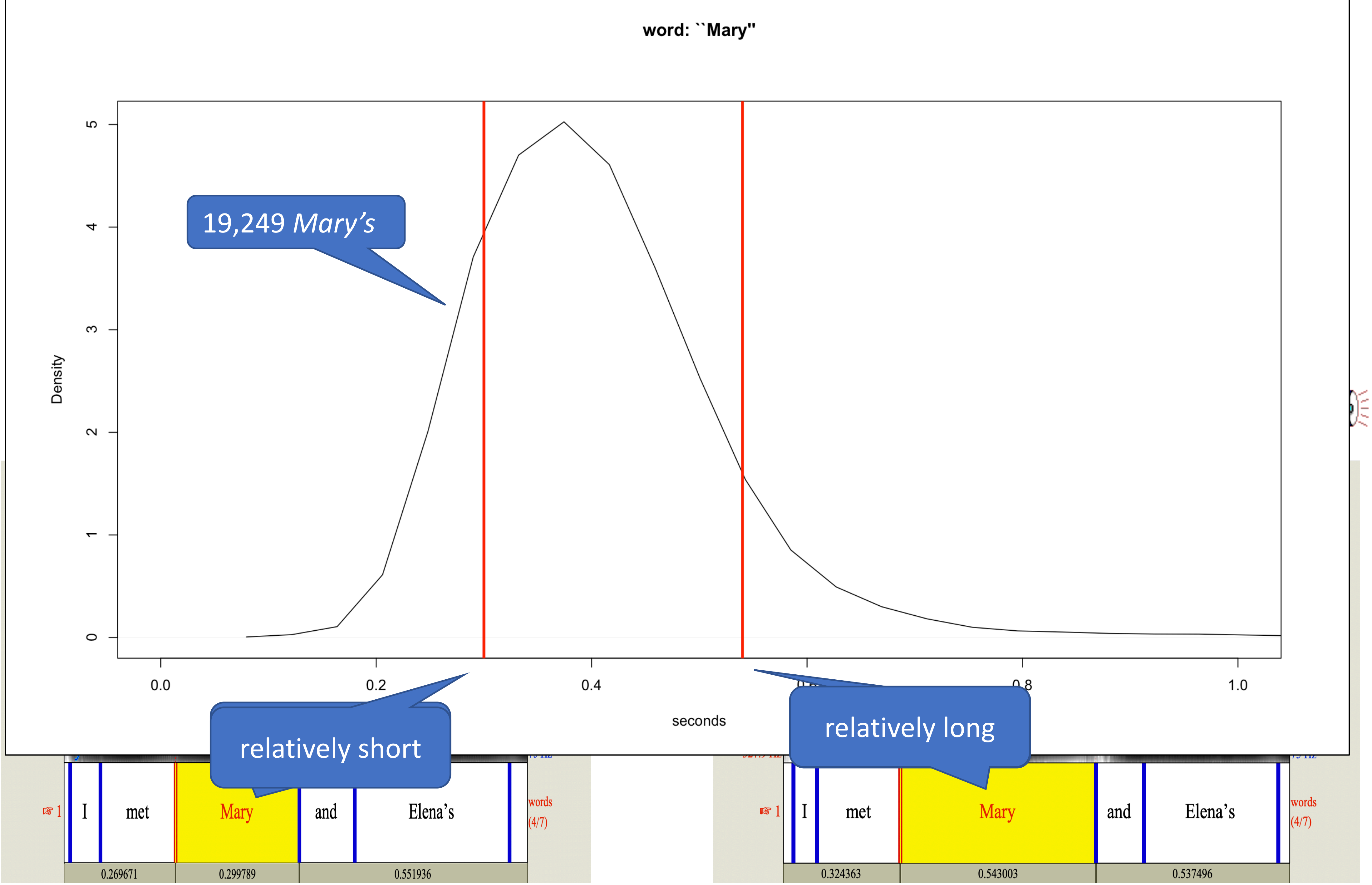
- I met [Mary and Elana]'s mother at the mall yesterday



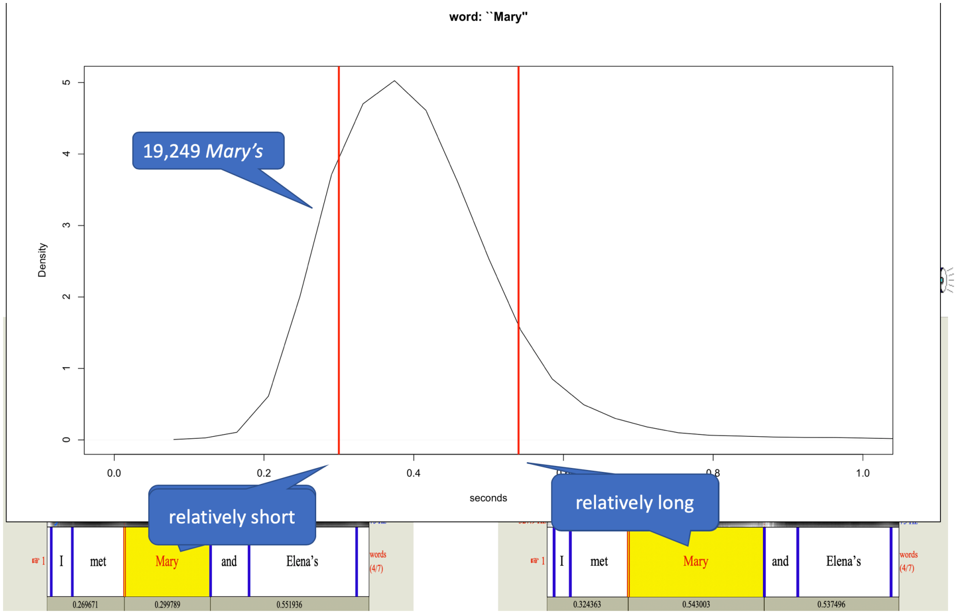
Conjunction: Wide Scope (0.5 sec)

- [I met Mary] and [Elana's mother] at the mall yesterday





word: ``Mary''
19,249 Mary's
Density
seconds
0.0
0.2
0.4
0.8
1.0
0
1
2
3
4
5
relatively short
relatively long
I
met
Mary
and
Elena's
words
(4/7)
0.269671
0.299789
0.551936
I
met
Mary
and
Elena's
words
(4/7)
0.324363
0.543003
0.537496

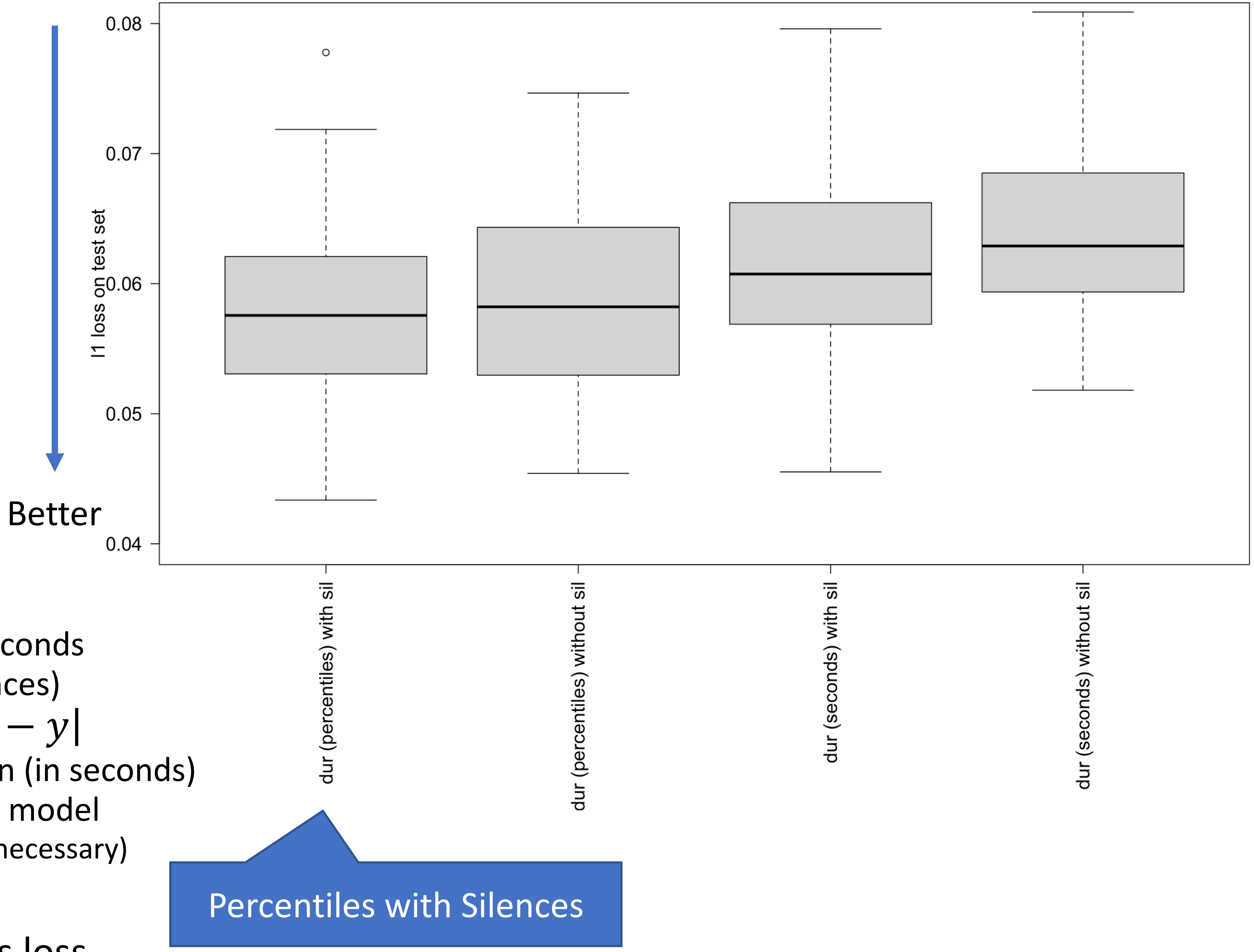


Percentile Transform

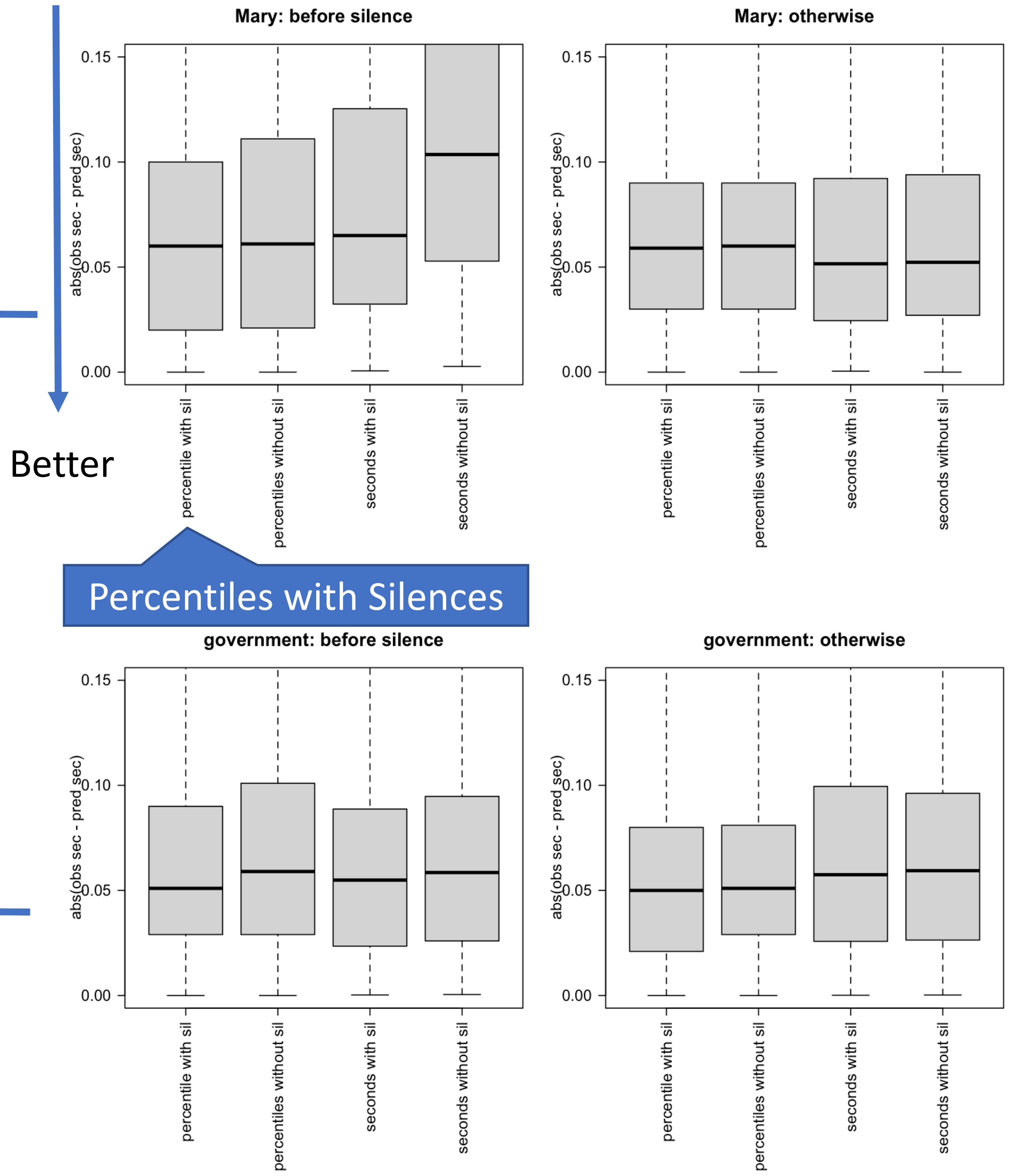
- Word Duration (seconds vs. percentiles)
  - Seconds (from forced alignments)
  - Percentiles:
    - based on durations of the same word in many other contexts
    - a definition of ``otherwise''
- Train:
  - Collect a large corpus of words (x) and durations (y)
  - Fine tune transformer (ERNIE/BERT) to predict $\hat{y}$ from x
- Inference
  - Input sequence of words (x); output sequence of predictions $(\hat{y})$
- Evaluation: Loss = $|\mathrm{sec}(word, \hat{y}) \ - y|$
  - where $\mathrm{sec}(word, \hat{y})$ converts prediction to seconds, if necessary
    - if prediction is already in seconds → do nothing (identity function)
    - if prediction is a percentile → invert the percentile transform

Evaluation

- Four conditions for training
  - duration:
    - measured in seconds
    - measured in percentiles
  - silences:
    - included in training
    - excluded from training
- Testing
  - Apples to apples
    - Convert all predictions to seconds
    - Evaluate on words (not silences)
  - For each token in test set $|\hat{y} - y|$
    - where y is observed duration (in seconds)
    - and $\hat{y}$ is the prediction from model
      - (converted to seconds, if necessary)
- Observations:
  - Percentile transform reduces loss
  - Ditto for silences (though less so)



Deep Dive: Mary + Government



Extensions

- Word durations depend on many factors
  - Word (type)
  - Context (other words near a particular mention), silences, phrasing
  - Emphasis/Accent/emotion
  - Speaker
  - Speaking Rate
- Percentile transform (and its inverse transform)
  - can be extended to depend not only on word and context
  - but many additional factors (conditioned on each audio book)

Unifying Themes

- ✓ Forced Alignment of Found Data
  - ✓ Input: Audio + Text
  - ✓ Output Timestamps: words, phones, silences
- ✓ Technologies
  - ✓ Machine Learning: Classification/Boosting/ERNIE/BERT
  - ✓ Fine-Tuning of language models with pauses (from audio)
    - ✓ Audio + Text are better together
- ✓ Linguistic Questions
  - ✓ Phrase final lengthening:
    - Some ``units'' are ``longer'' than ``otherwise'' in certain ``contexts''
  - ➢ **t/d deletion**
    - **Some ``units'' are ``deleted'' in certain ``contexts''**
- Practical Questions
  - Dementia Challenge: Distinguish AD from controls
  - Observation: disfluencies are often associated with pauses

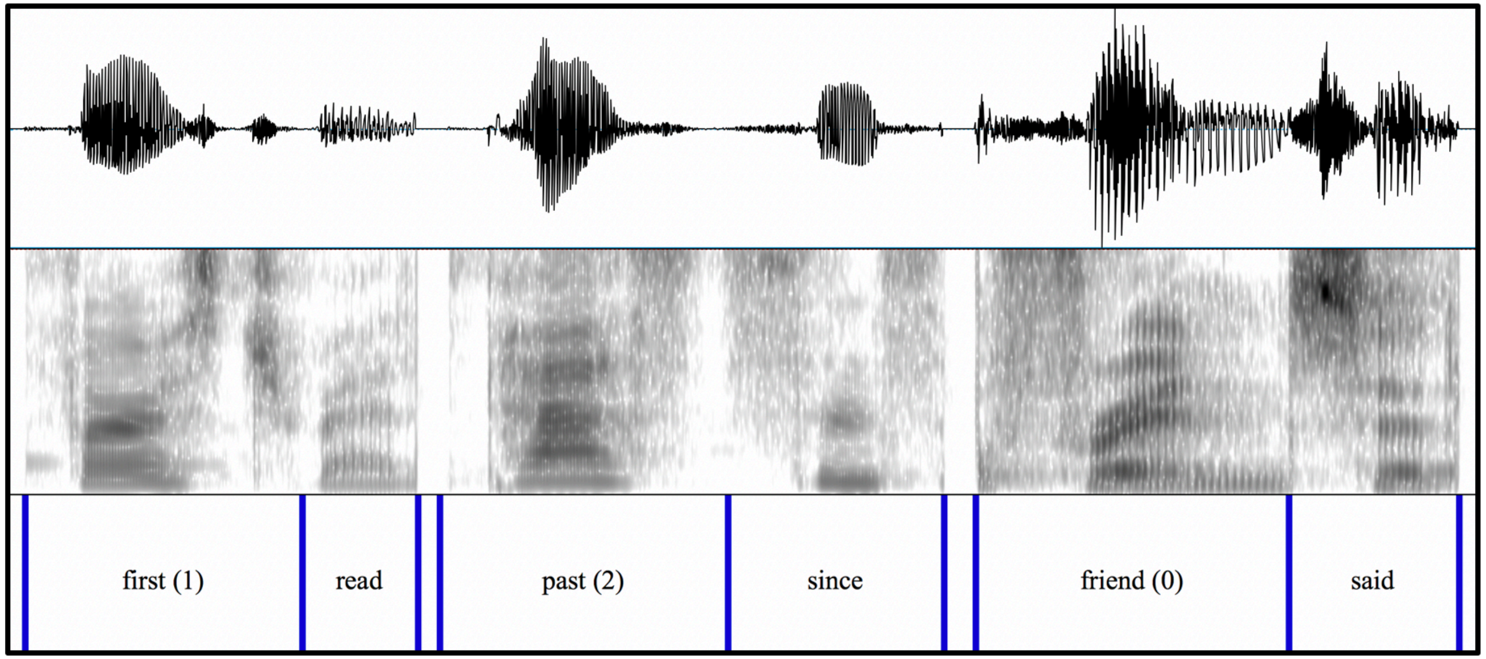
Found Data	Size (M words)
Audio Books	111.4
SCOTUS	70.0
Audio BNC	7.1
Tedlium	5.7
History	5.0
Presidential	1.5
CommonVoice	0.7

Detection and analysis of T/D deletion in Librispeech

t/d deletion

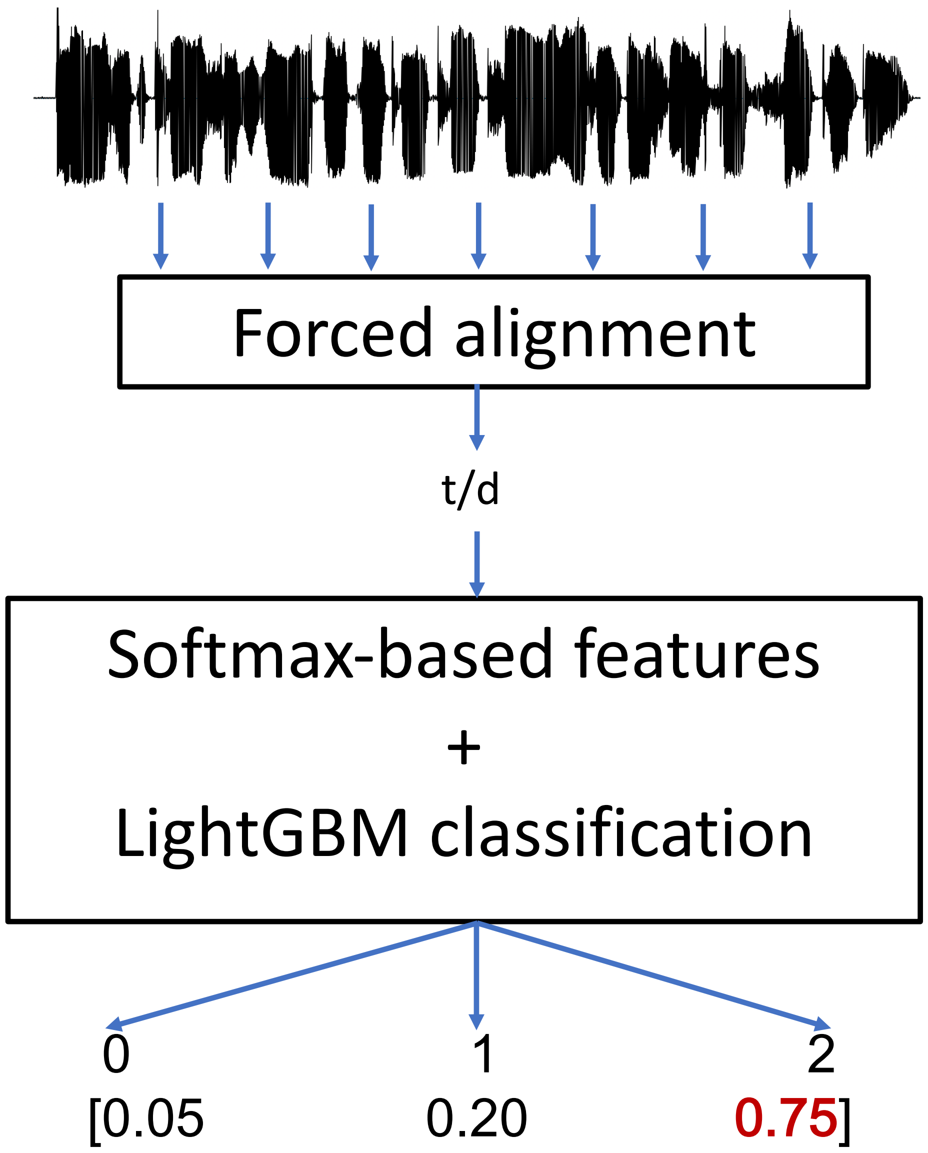
- Categorical?

  0 (deletion), **1** (full realization), **2** (partial realization)



- Manual annotation on t/d deletion (binary): 80% agreement

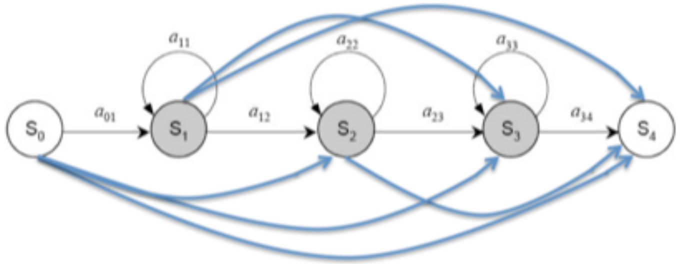
Automatic identification (1)



Automatic identification (2)

- Step 1: Forced alignment

  - Skip-state HMMs for word-final /t/ and /d/



  - Which can identify t/d deletion with 79.1% accuracy on TIMIT

    - duration = 0 ↔ t/d deletion

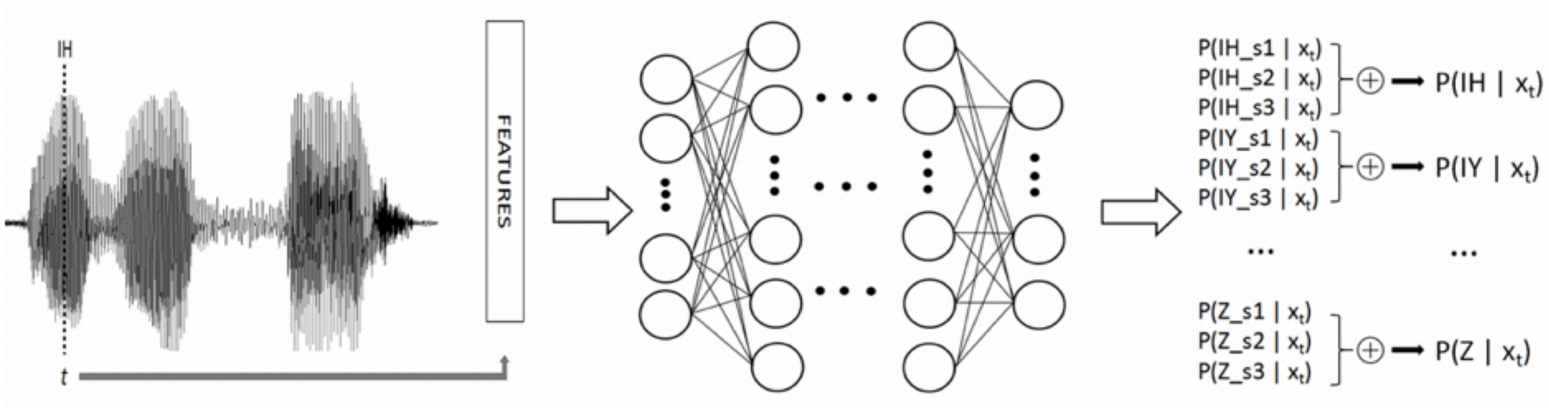
    - better than using alternative pronunciations (73.6%)
      best /B EH1 S T/
      best /B EH1 S/

Automatic identification (3)

- Step 2: Extract features

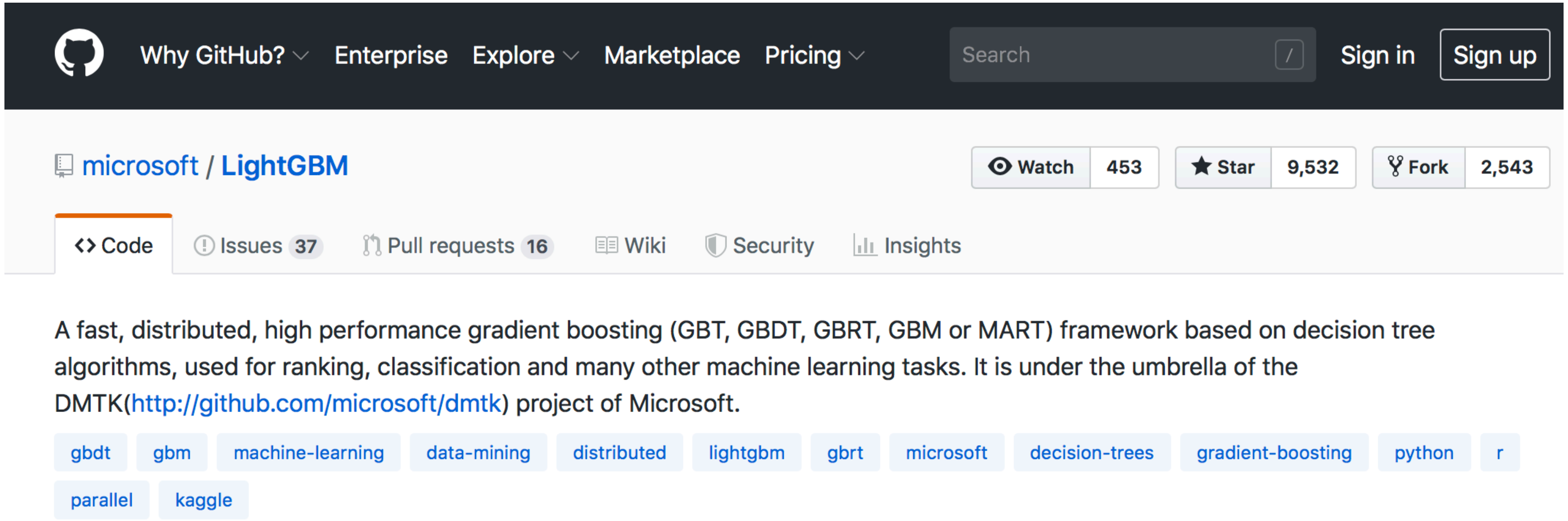
  - At Three points
    - Onset, center, offset
    - Three times at the same position if duration = 0

  - Softmax-based features
    - Kaldi/TDNN trained on Librispeech
    - 70-dim feature vector: 69 phonemes (with stress) + sil



Automatic identification (4)

- Step 3: LightGBM classification
  - Combine decision trees (weak learners) to minimize the loss function (gradient boosting).



Automatic identification (5)

- Evaluation
  - TIMIT
    - Based on phone transcription
    - Labels: 0 (no transcription), 1 (/t, d, dx, jh/), and 2 (/ tcl, dcl, q/).
  - Librispeech
    - Manually annotated 1,800 tokens
    - Labels: 0, 1, 2
  - Accuracy of two-class classification
    - Deletion (0); No deletion (1,2)

	Forced alignment (skip-state HMMs)	LightGBM after forced alignment
TIMIT	79.1%	93.7%
Librispeech	80.6%	86.7%

Large scale analysis (1)

- Data: Librispeech
  - excluding:
    - Uncommon words (frequency < 100)
    - The word “and” (frequency > 300,000)
  - word-final t/d preceded by a consonant
  - 502,481 tokens, 818 word types
- Classification
  - Forced aligner and TDNN were trained on entire Librispeech
  - LightGBM was trained on manually annotated Librispeech data

Large scale analysis (2)

- Statistical significance

  - Logistic regression
    Six main factors: t/d, preceding phone, following phone, morphological class, word frequency, PND

  - All main factors except word frequency have a significant effect.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.073306	0.089026	-23.289	< 2e-16 ***
T	-0.275848	0.064475	-4.278	1.88e-05 ***
p2	-1.616205	0.073061	-22.121	< 2e-16 ***
f2	1.721275	0.091135	18.887	< 2e-16 ***
f3	2.543745	0.062242	40.868	< 2e-16 ***
c2	-1.290159	0.182971	-7.051	1.77e-12 ***
c3	-0.799413	0.064708	-12.354	< 2e-16 ***
frequency	0.038280	0.028446	1.346	0.178391
density	-0.582366	0.068577	-8.492	< 2e-16 ***

Conclusions

- We developed a new method for automatic identification of t/d deletion in continuous speech. Our method achieved 93.7% accuracy on TIMIT and 86.7% on human-annotated data from Librispeech.

- A large scale analysis on Librispeech showed that word frequency was not a significant factor in determining the rate of t/d deletion, although the interactions between word frequency and other factors were significant.

- Phonological Neighborhood Density showed a much stronger effect on t/d deletion than word frequency. t/d is less likely to be deleted when PND is higher (i.e., having more neighbors).

- Our results on the effects of phonological and morphological factors are largely consistent with previous studies.

Unifying Themes

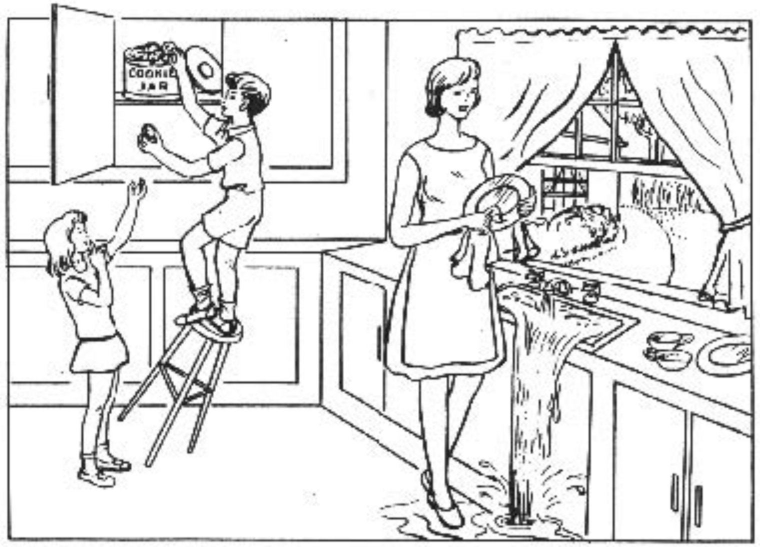
- ✓ Forced Alignment of Found Data
  - ✓ Input: Audio + Text
  - ✓ Output Timestamps: words, phones, silences
- ✓ Technologies
  - ✓ Machine Learning: Classification/Boosting/ERNIE/BERT
  - ✓ Fine-Tuning of language models with pauses (from audio)
    - ✓ Audio + Text are better together
- ✓ Linguistic Questions
  - ✓ Phrase final lengthening:
    - Some ``units'' are ``longer'' than ``otherwise'' in certain ``contexts''
  - ✓ t/d deletion
    - ✓ Some ``units'' are ``deleted'' in certain ``contexts''
- **Practical Questions**
  - ➢ **Dementia Challenge: Distinguish AD from controls**
  - **Observation: disfluencies are often associated with pauses**

Found Data	Size (M words)
Audio Books	111.4
SCOTUS	70.0
Audio BNC	7.1
Tedlium	5.7
History	5.0
Presidential	1.5
CommonVoice	0.7

Detection of Alzheimer’s disease

The ADReSS Challenge

- Alzheimer's Dementia Recognition through Spontaneous Speech
- Dataset:
  - Training: 108 recordings + transcripts; 54 control + 54 ad
  - Test: 48 recordings + transcripts



- Tasks:
  - A binary classification of AD and non-AD
  - To predict scores of Mini-Mental State Examination (MMSE)

Challenges Awards: ADReSS

Laudator: *Saturnino Luz*

Interspeech 2020

ADReSS

The ADReSS Alzheimer's Dementia Recognition Challenge

Award Winner:

Jiahong Yuan, Yuchen Bian, Xingyu Cai,
Jiaji Huang, Zheng Ye and Kenneth Church
(Baidu Research USA)

"Disfluencies and Fine-Tuning Pre-trained Language Models for Detection of Alzheimer's Disease"

Organizers:
Saturnino Luz, Fasih Haider, Sofia de la Fuente,
Davida Fromm, Brian MacWhinney

Session Chairs:
Isabel Trancoso, Nick Campbell

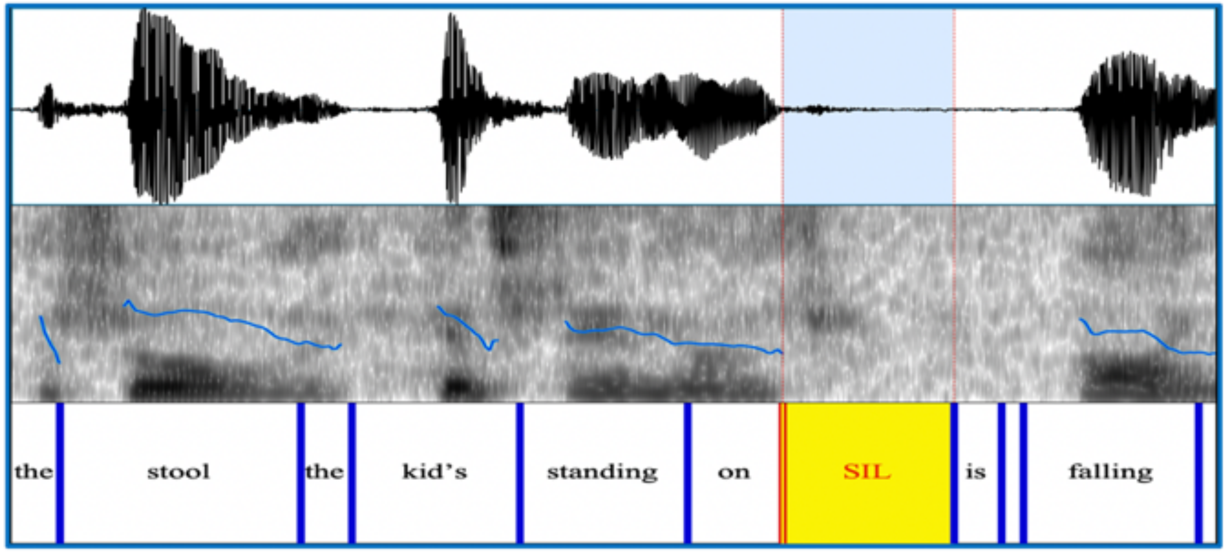
Classification method and experiments

- Step 1: Forced alignment and pause encoding
- Step 2: Fine-tuning BERT/ERNIE using pause-inserted text
- Step 3: Ensemble over many runs of fine-tuning

Forced alignment and pause encoding

Input: transcript + audio

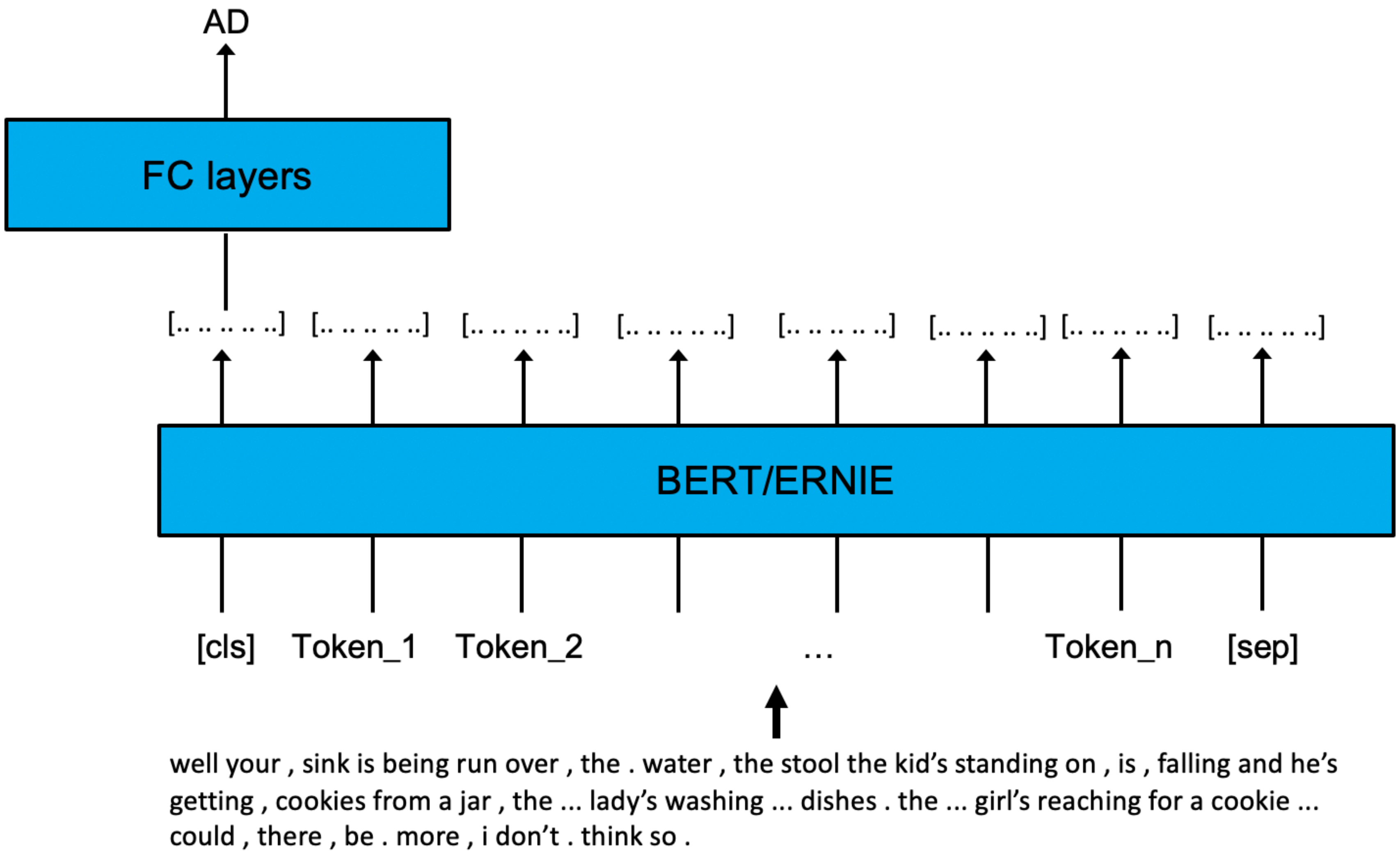
Forced alignment



Pause encoding (3p): <0.5s (,); 0.5-2s (.); >2s (...)

Output: well your , sink is being run over , the . water , the stool the kid's standing on , is , falling and he's getting , cookies from a jar , the ... lady's washing ... dishes . the ... girl's reaching for a cookie ... could , there , be . more , i don't . think so .

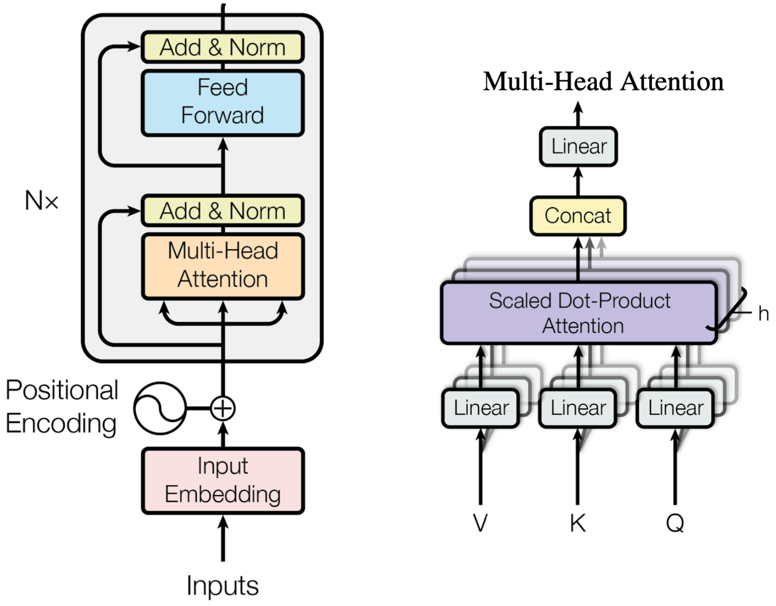
Fine-tuning BERT/ERNIE for AD classification



BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

- Using multi-head self-attention to capture associations among words.
  - has more than **100M** parameters
  - pretrained on **billions** of words (wikipedia, bookcorpus, etc.)



"Attention Is All You Need"

Results and conclusions

- Evaluation on the test set (majority vote of 35 runs):

	Precision		Recall		F1		Acc
	non-AD	AD	non-AD	AD	non-AD	AD	
Baseline[6]	0.700	0.830	0.870	0.620	0.780	0.710	0.750
BERT0p	0.742	0.941	0.958	0.667	0.836	0.781	0.813
BERT3p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
BERT6p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
ERNIE0p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
ERNIE3p	0.852	0.952	0.958	0.833	0.902	0.889	**0.896**

1. Disfluencies and language problems in Alzheimer’s Disease can be naturally modeled by fine-tuning Transformer-based pre-trained language models.

2. The best accuracy was obtained with ERNIE, plus an encoding of pauses.

3. We found that *um* was used much less frequently in Alzheimer’s speech.

Unifying Themes

- ✓ Forced Alignment of Found Data
  - ✓ Input: Audio + Text
  - ✓ Output Timestamps: words, phones, silences
- ✓ Technologies
  - ✓ Machine Learning: Classification/Boosting/ERNIE/BERT
  - ✓ Fine-Tuning of language models with pauses (from audio)
    - ✓ Audio + Text are better together
- ✓ Linguistic Questions
  - ✓ Phrase final lengthening:
    - ✓ Some ``units'' are ``longer'' than ``otherwise'' in certain ``contexts''
  - ✓ t/d deletion
    - ✓ Some ``units'' are ``deleted'' in certain ``contexts''
- ✓ Practical Questions
  - ✓ Dementia Challenge: Distinguish AD from controls
  - ✓ Observation: disfluencies are often associated with pauses

Found Data	**Size (M words)**
Audio Books	111.4
SCOTUS	70.0
Audio BNC	7.1
Tedlium	5.7
History	5.0
Presidential	1.5
CommonVoice	0.7