Automatic detection of sociolinguistic variation in forced alignment

George Bailey
University of Manchester

NWAV44 - 24 October 2015

1. Introduction

Research questions

Forced alignment

Hidden Markov Models

Pronouncing dictionary

2. Methodology

Dictionary ‘hacking’

Measuring accuracy

3. Results

Overview

Detailed analysis

Rate of speech

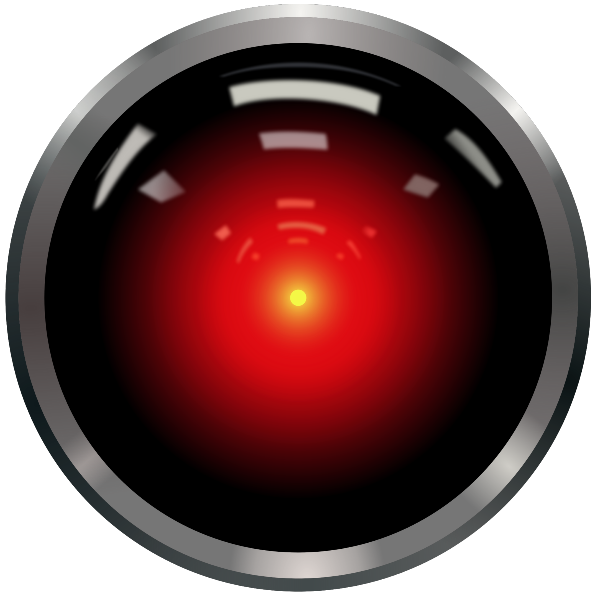
4. Conclusion

Research questions

- To investigate the possibility of using forced alignment to automatically code phonological variation
- To assess the accuracy and reliability of this methodology
- To provide insight into the patterning of its errors

Why?

- Increased efficiency, with one fewer step in the data-collection workflow
- Particularly important given the ‘big data’ trend
  - Use of FAVE-extract for automatic formant measurements, e.g. 3000-9000 vowel measurements per interview in the PNC (Labov et al. 2013)
  - Emergence of aligners like DARLA (Reddy & Stanford 2015) that remove the need for transcription entirely
- Arguably more reliable
  - less prone to human error
  - more replicable



1. Introduction

Research questions

Forced alignment

Hidden Markov Models

Pronouncing dictionary

2. Methodology

Dictionary ‘hacking’

Measuring accuracy

3. Results

Overview

Detailed analysis

Rate of speech

4. Conclusion

Forced alignment

- Discussion here will focus on FAVE - the University of Pennsylvania's 'Forced Alignment and Vowel Extraction' suite (Rosenfelder et al. 2014)
- Other aligners (e.g. PLA, Gorman et al. 2011) are available!
- Mechanisms and output of forced-alignment largely consistent across different suites

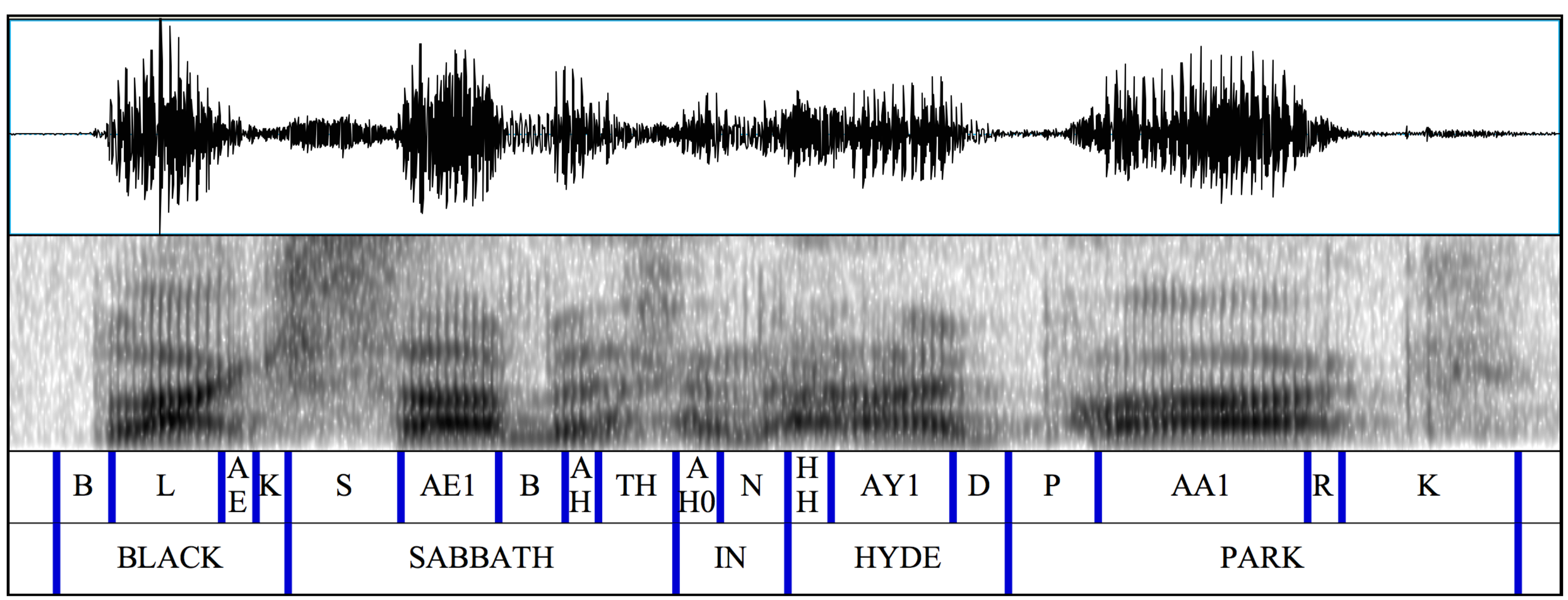
Forced alignment

What does it do?

Input: Audio + word-level, orthographic transcription

Forced Alignment

Output: Time-aligned Praat TextGrid with phone- and word-level tiers



Forced alignment

How does it do it?

- By comparing the speech signal with pre-established acoustic models
- By making reference to a standard pronouncing dictionary

1. Introduction

Research questions

Forced alignment

Hidden Markov Models

Pronouncing dictionary

2. Methodology

Dictionary 'hacking'

Measuring accuracy

3. Results

Overview

Detailed analysis

Rate of speech

4. Conclusion

Hidden Markov Models

- Hidden Markov Model Toolkit (HTK) - natural language processor (see Ghahramani 2001)
- FAVE's acoustic models are based on American English, trained on the SCOTUS corpus
    - still performs well on British English data (see MacKenzie & Turton 2013)

1. Introduction

- Research questions
- Forced alignment
  - Hidden Markov Models
  - Pronouncing dictionary

2. Methodology

- Dictionary ‘hacking’
- Measuring accuracy

3. Results

- Overview
- Detailed analysis
- Rate of speech

4. Conclusion

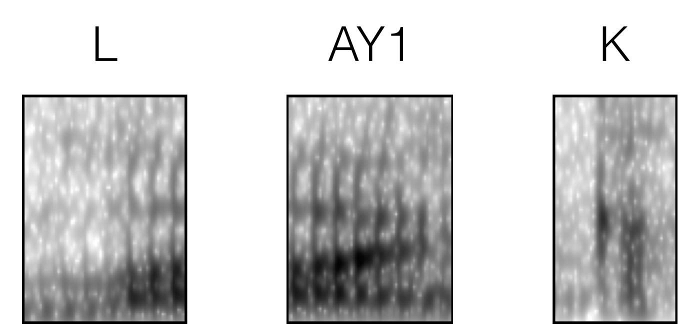
Pronouncing dictionaries

- Pronouncing dictionaries provide phone-level transcriptions (in Arpabet) for a particular language's lexicon

- FAVE uses the Carnegie Mellon University dictionary (CMUdict) based on General American orthography and phonology

    - wide coverage of lexicon with over 134,000 entries

`everyday is` **`like`** `sunday` → `LIKE: L AY1 K` →



1. Introduction

Research questions

Forced alignment

Hidden Markov Models

Pronouncing dictionary

2. Methodology

Dictionary ‘hacking’

Measuring accuracy

3. Results

Overview

Detailed analysis

Rate of speech

4. Conclusion

Dictionary ‘hacking’

- Crucially, these dictionaries provide only broad, phonemic transcriptions

- They *can* contain multiple entries for the same word

  - e.g. *present* - P R **EH1** Z **AH0** N T
P R **AH0** Z **EH1** N T

- What happens when the aligner encounters a word with multiple possible realisations?

  - It compares the output probabilities from all potential models and picks the best-fitting one

Dictionary ‘hacking’

- This is the methodology employed here with sociolinguistic variables
- Expansion of the pronouncing dictionary to represent the surface output from phonological processes
- Comparable to Yuan & Liberman (2011) and Milne (2014)

Dictionary ‘hacking’

- Variables:
  - (td)-deletion /t, d/ $\longrightarrow$ ø
    J AH1 S **T**
    J AH1 S
  - (th)-fronting /θ, ð/ $\longrightarrow$ [f, v]
    N AO1 **TH**
    N AO1 **F**
  - (h)-dropping /h/ $\longrightarrow$ ø
    H EY1 T
    EY1 T
- Python scripts were used to identify words that fall within each variable’s envelope of variation
  - Addition of 8371 (td), 3483 (th) and 5302 (h) entries

1. Introduction

- Research questions
- Forced alignment
  - Hidden Markov Models
  - Pronouncing dictionary

2. Methodology

- Dictionary ‘hacking’
- Measuring accuracy

3. Results

- Overview
- Detailed analysis
- Rate of speech

4. Conclusion

Measuring accuracy

- Hour-long sociolinguistic interview with a 20 year-old female speaker from Manchester, England - sampling rate of 44,100 Hz
    - 249 tokens of (h), 293 of (td), and 364 of (th)
- Alignment carried out using the expanded pronouncing dictionaries
- FAVE's discriminative judgements compared to manually-coded human judgements
    - Two measures: percentage agreement and Cohen's Kappa (see Carletta 1966)
- Second round of manual coding carried out by another human transcriber to establish inter-transcriber agreement rates

1. Introduction

Research questions

Forced alignment

Hidden Markov Models

Pronouncing dictionary

2. Methodology

Dictionary 'hacking'

Measuring accuracy

3. Results

Overview

Detailed analysis

Rate of speech

4. Conclusion

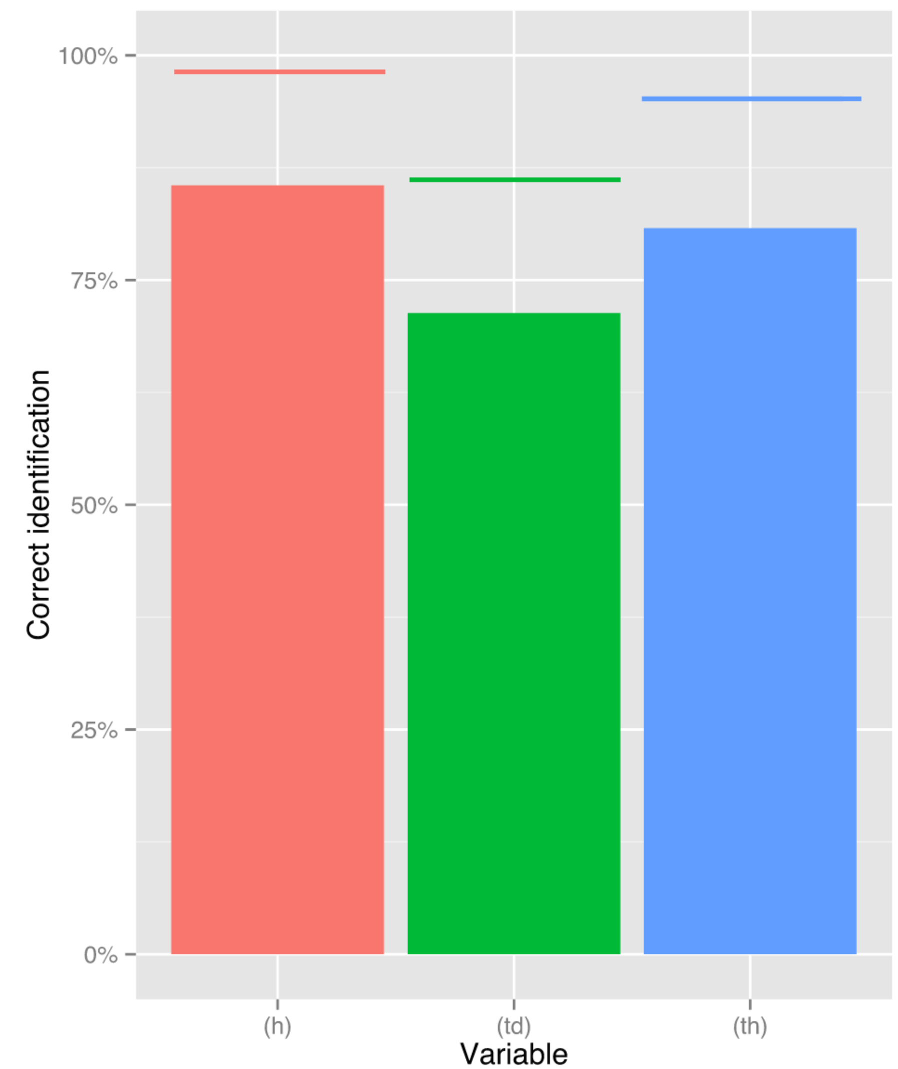
Results

Overview

	FAVE agreement		Inter-transcriber agreement		N
	%	κ	%	κ	
(h)	85.54%	0.63	97.19%	0.91	249
(td)	71.33%	0.43	84.98%	0.70	293
(th)	79.67%	0.57	92.58%	0.81	364
TOTAL:	**78.59%**	**0.55**	**91.39%**	**0.81**	**906**

- “Moderate” FAVE-agreement
- “Almost perfect” inter-transcriber agreement

21.4	15.9	25.4	χ^2
<0.01	<0.01	<0.01	p
*	*	*	



Results

- Important to perform detailed analysis of FAVE's ability to recognise both application and *non*-application of these variables
- As such, FAVE's discriminative judgements are classified into four categories:
  - true positives - correct identification of application
  - true negatives - correct identification of non-application
  - false positives - incorrect identification of application (≈ type I error)
  - false negatives - incorrect identification of non-application (≈ type II error)

(h)

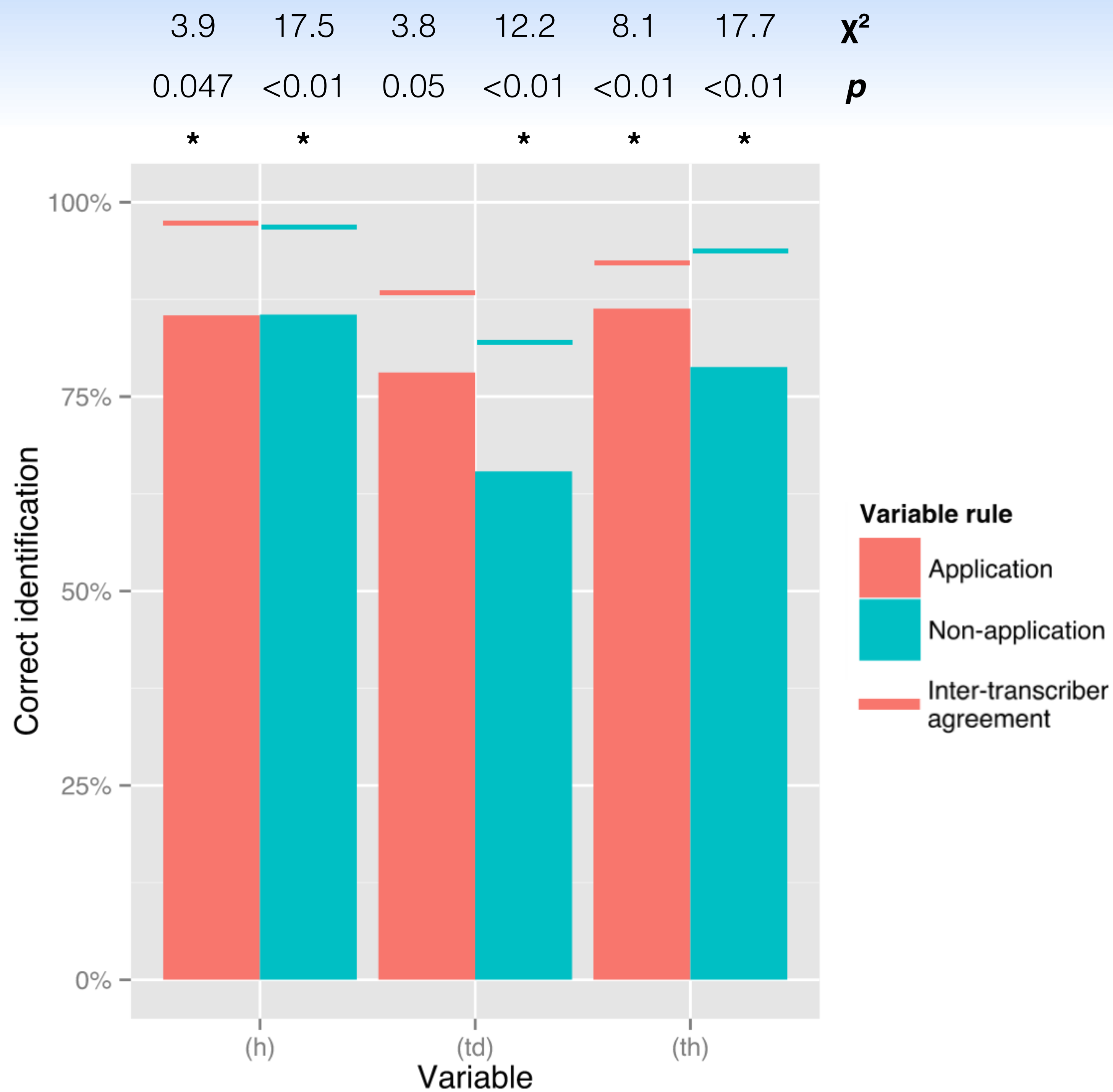
		Human ø	Human [h]
FAVE	ø	47 **85.5%**	28 **14.4%**
FAVE	[h]	8 **14.5%**	166 **85.6%**

(td)

		ø	[t, d]
FAVE	ø	107 **78.1%**	54 **34.6%**
FAVE	[t, d]	30 **21.9%**	102 **65.4%**

(th)

		[f, v]	[θ, ð]
FAVE	[f, v]	82 **86.3%**	57 **21.2%**
FAVE	[θ, ð]	13 **13.7%**	212 **78.8%**



- Lower accuracy for (td) can be attributed to non-application
- Inter-transcriber agreement suffers comparably

Results

- Also important to consider voiced and voiceless segments separately

- Especially when the distribution isn't equal:

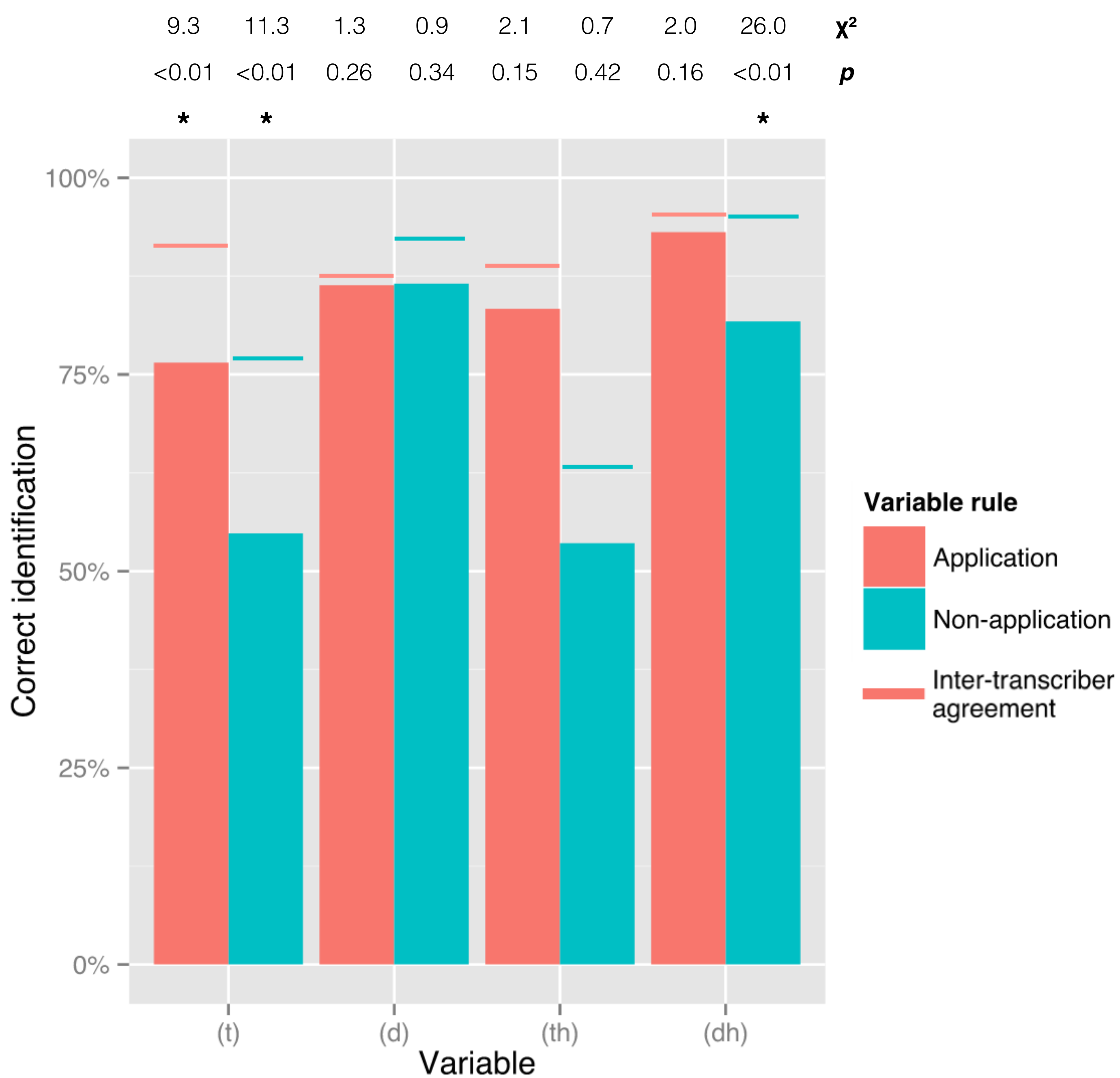
    - 204 tokens of (t) ~ 71 tokens of (d)

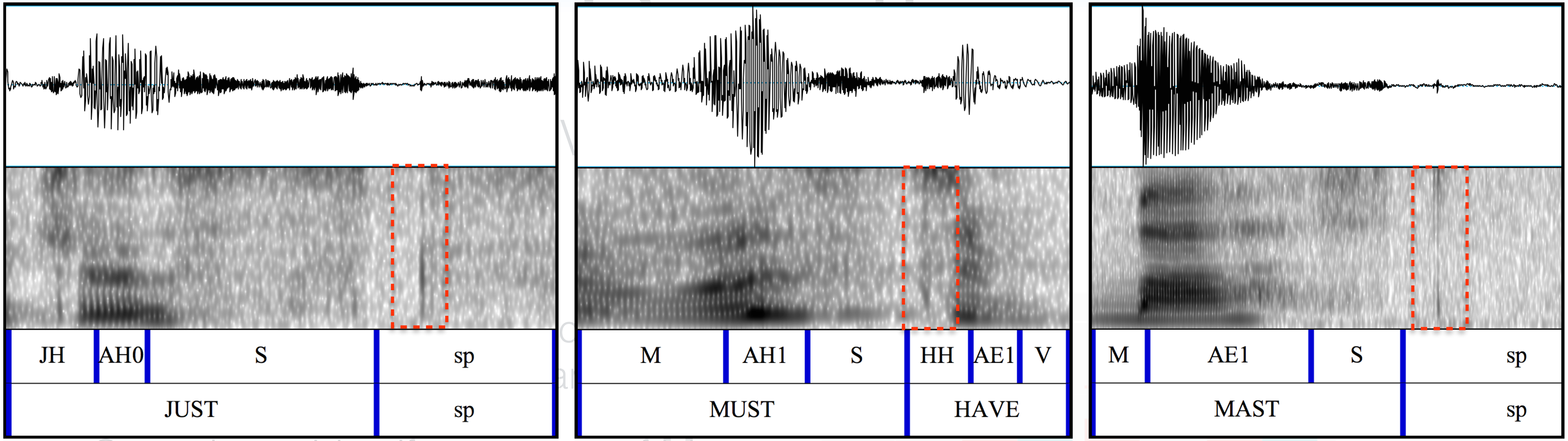
    - 90 tokens of (th) ~ 235 tokens of (dh)

Results

Voiced vs. voiceless

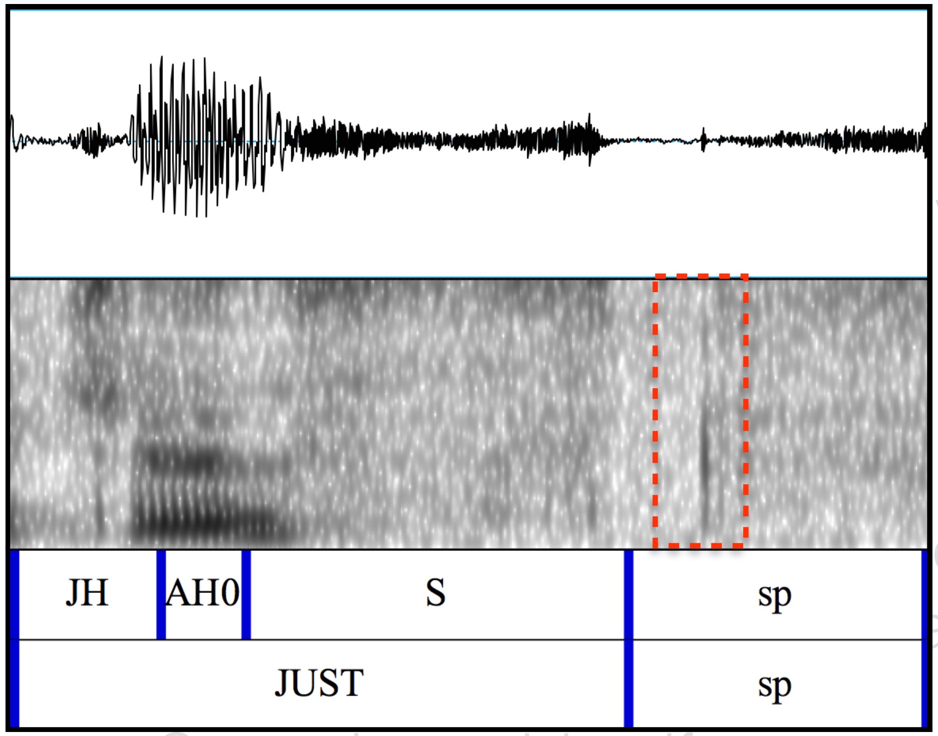
- Lowest accuracy for non-application on the voiceless segments /t/ and /θ/
  - Struggles to identify presence of [t]
  - Misidentifies [θ] as [f]
- Lenited quality of word-final /t/ makes it hard to identify?

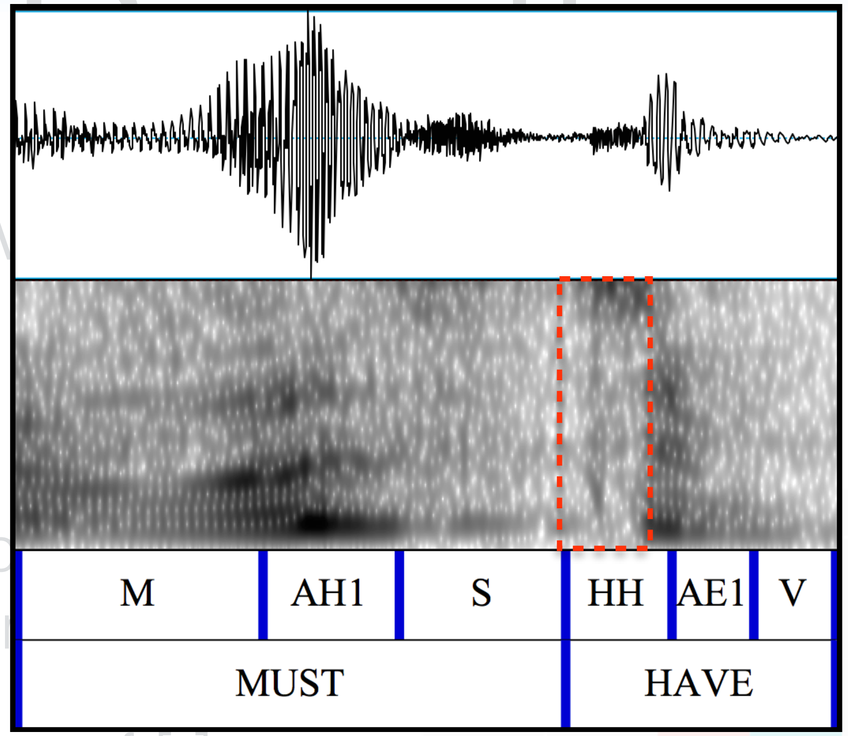


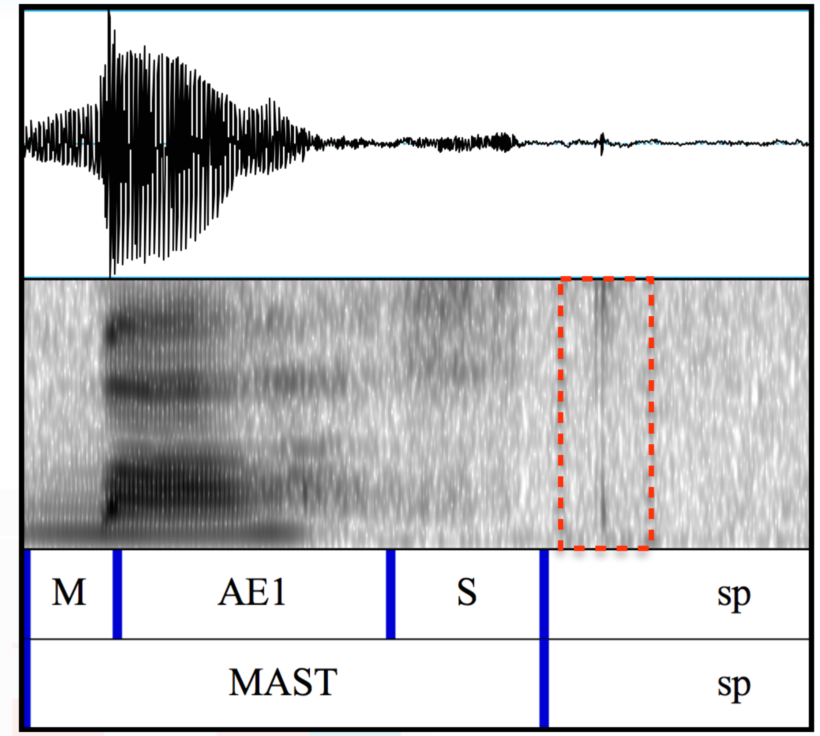


- Struggles to identify presence of [t]
- Misidentifies [θ] as [f]

- Lenited quality of word-final /t/ makes it hard to identify?

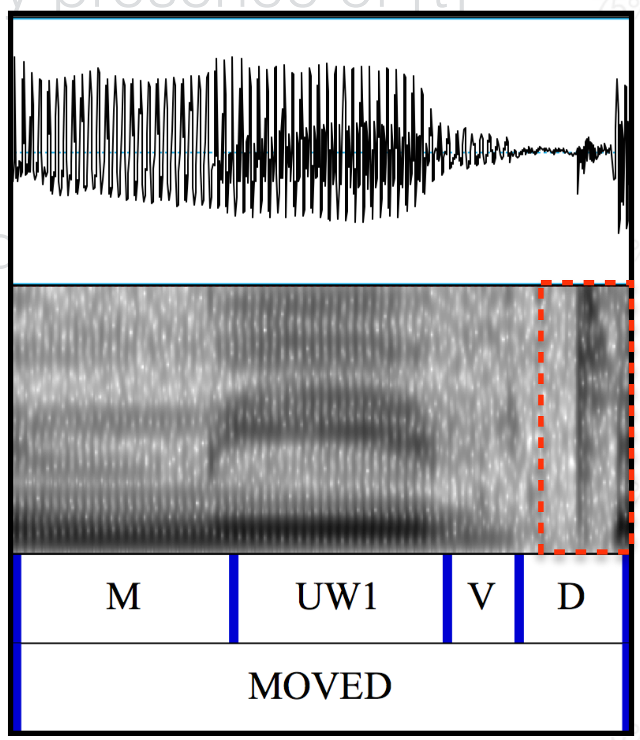


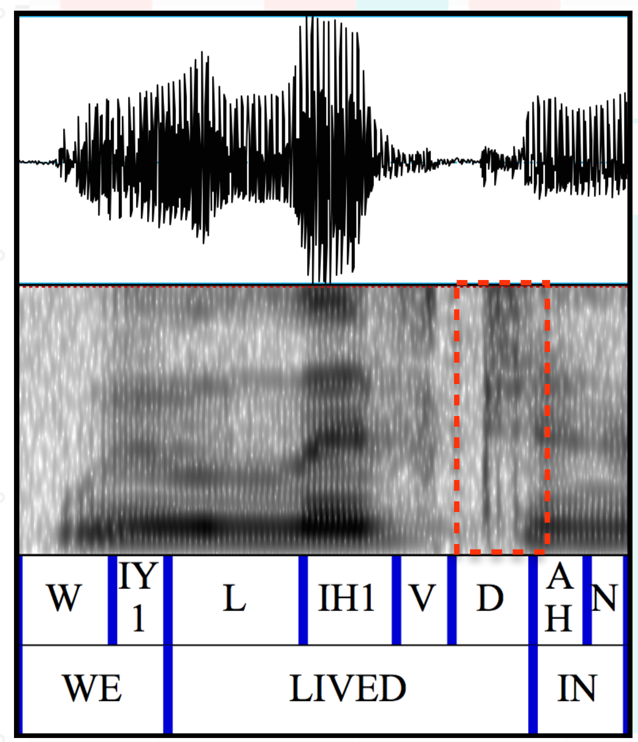




- Struggles to identify presence of [t]
- Misidentifies [θ] as

- Lenited quality of word
  it hard to identify?

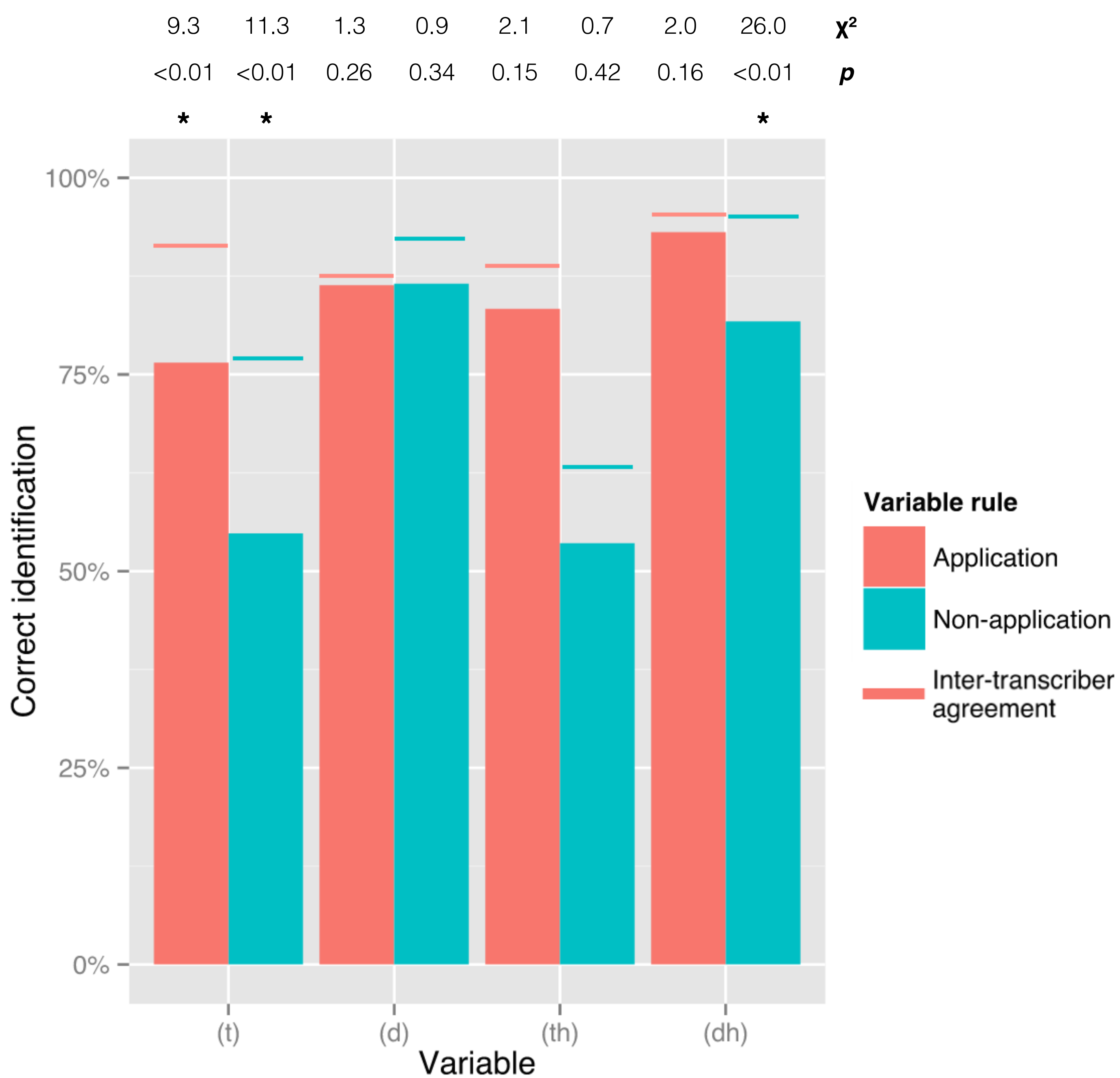




Results

Voiced vs. voiceless

- Lowest accuracy for non-application on the voiceless segments /t/ and /θ/
  - Struggles to identify presence of [t]
  - Misidentifies [θ] as [f]
- Lenited quality of word-final /t/ makes it hard to identify?
- Over-zealous in seeking out [f]?
- Once again, inter-transcriber agreement sees similar drops for these segments



1. Introduction

Research questions

Forced alignment

Hidden Markov Models

Pronouncing dictionary

2. Methodology

Dictionary ‘hacking’

Measuring accuracy

3. Results

Overview

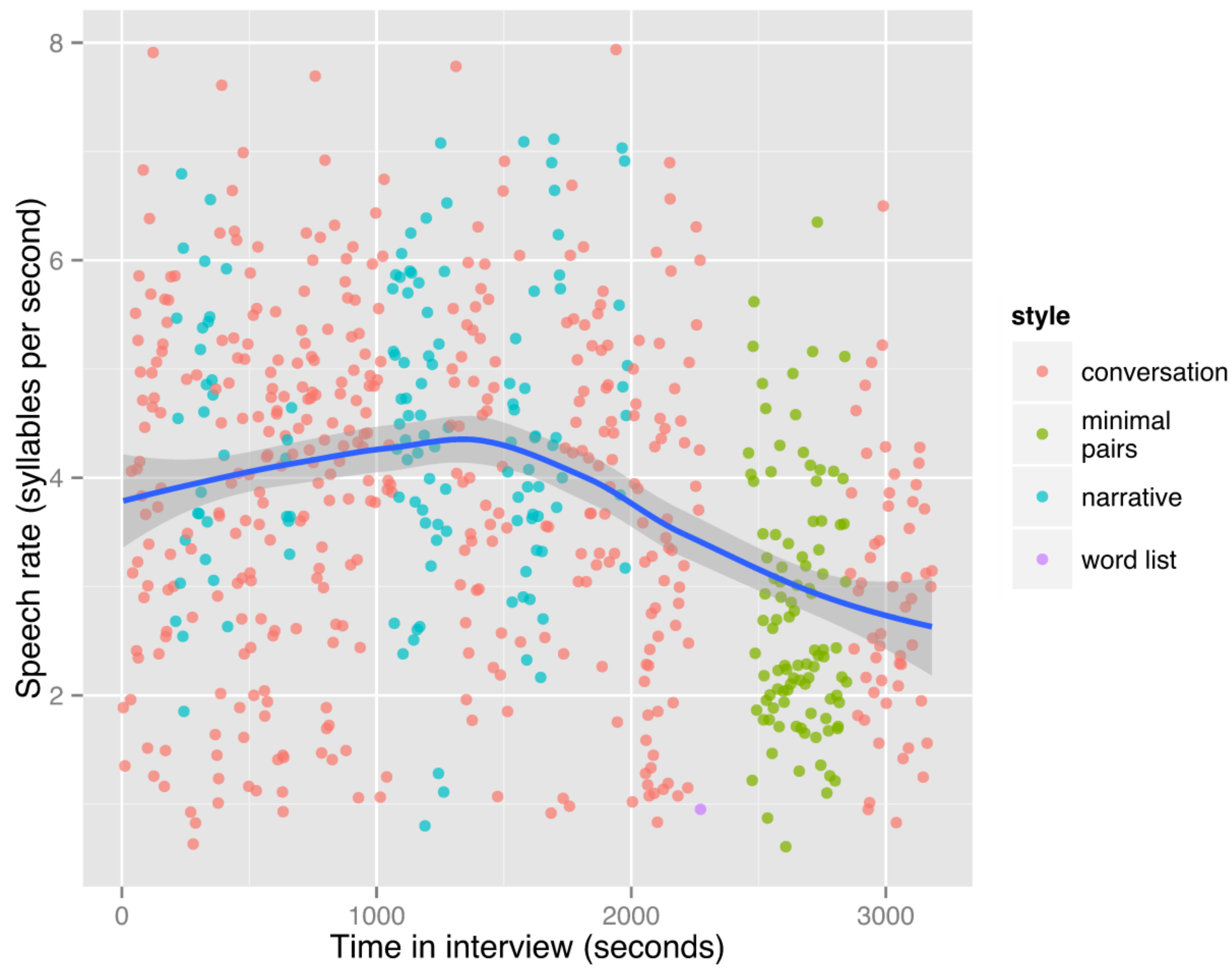
Detailed analysis

Rate of speech

4. Conclusion

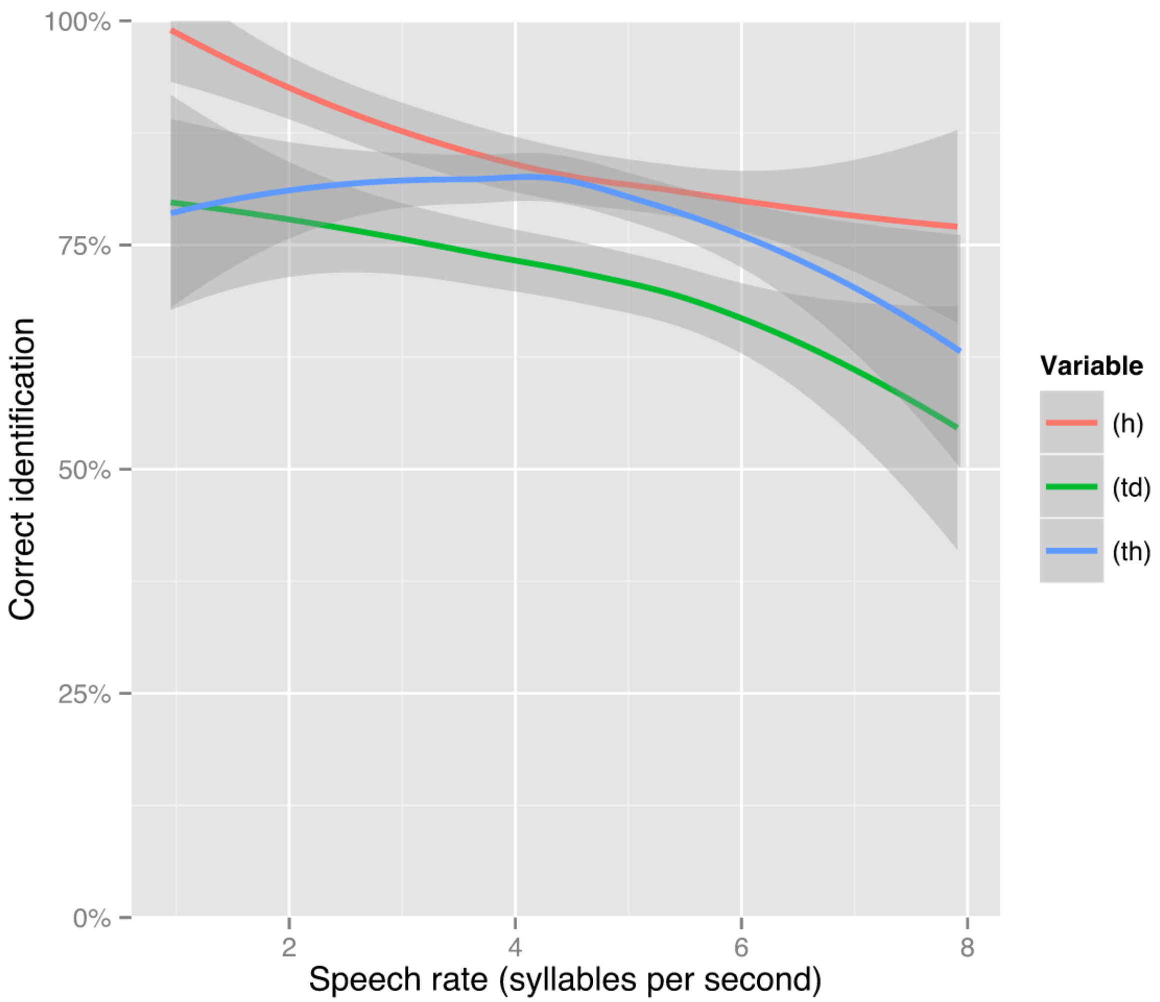
Rate of speech

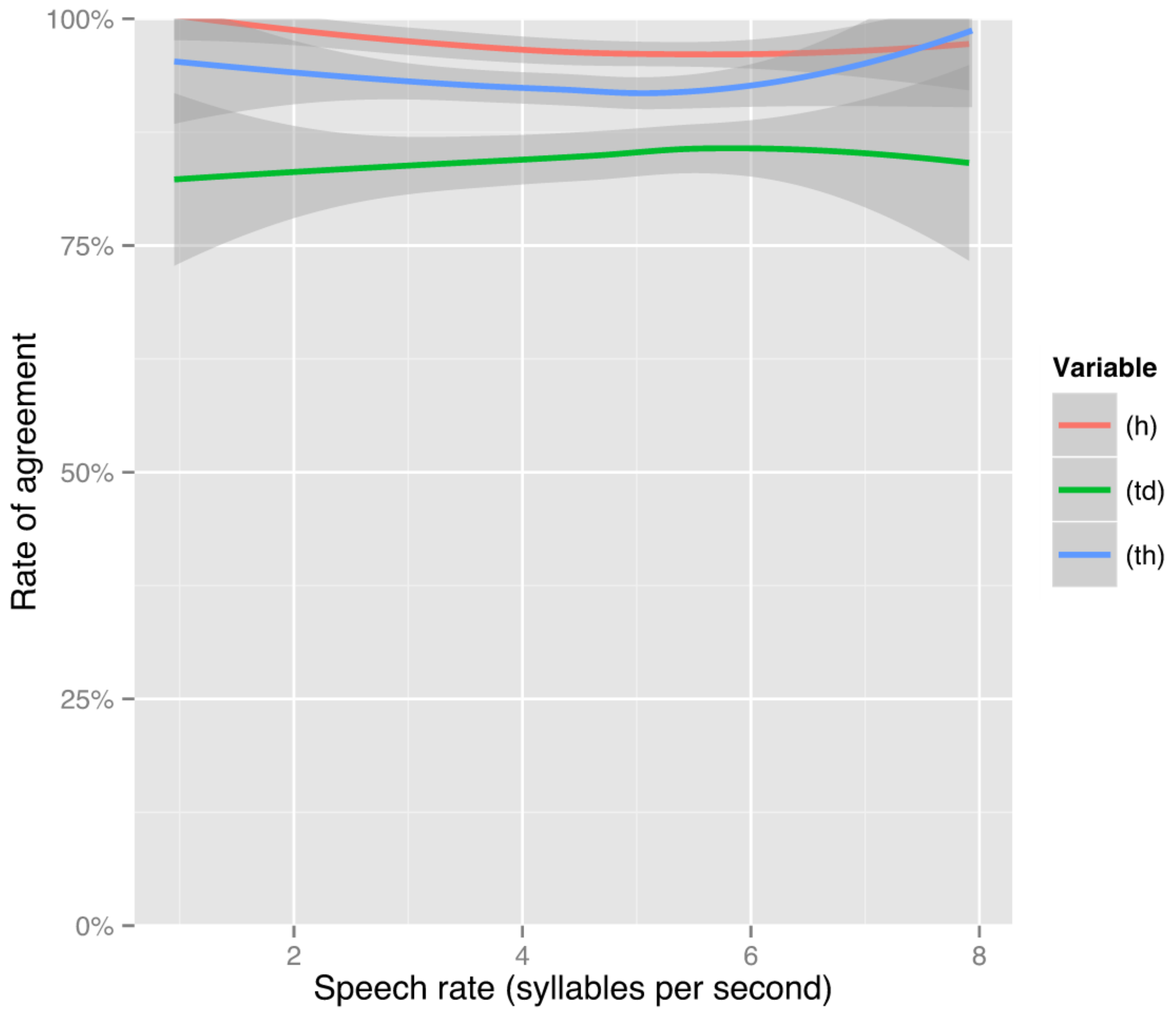
- Speech rate can vary dramatically throughout a sociolinguistic interview, often corresponding with changes in formality
    - e.g. narratives of personal experience = fastest
    - e.g. word lists = slowest
- Narrative = 4.35 sylls per/s
- Conversation = 3.69 sylls per/s
- Minimal pairs = 2.71 sylls per/s
- Word list = 0.95 sylls per/s



Rate of speech

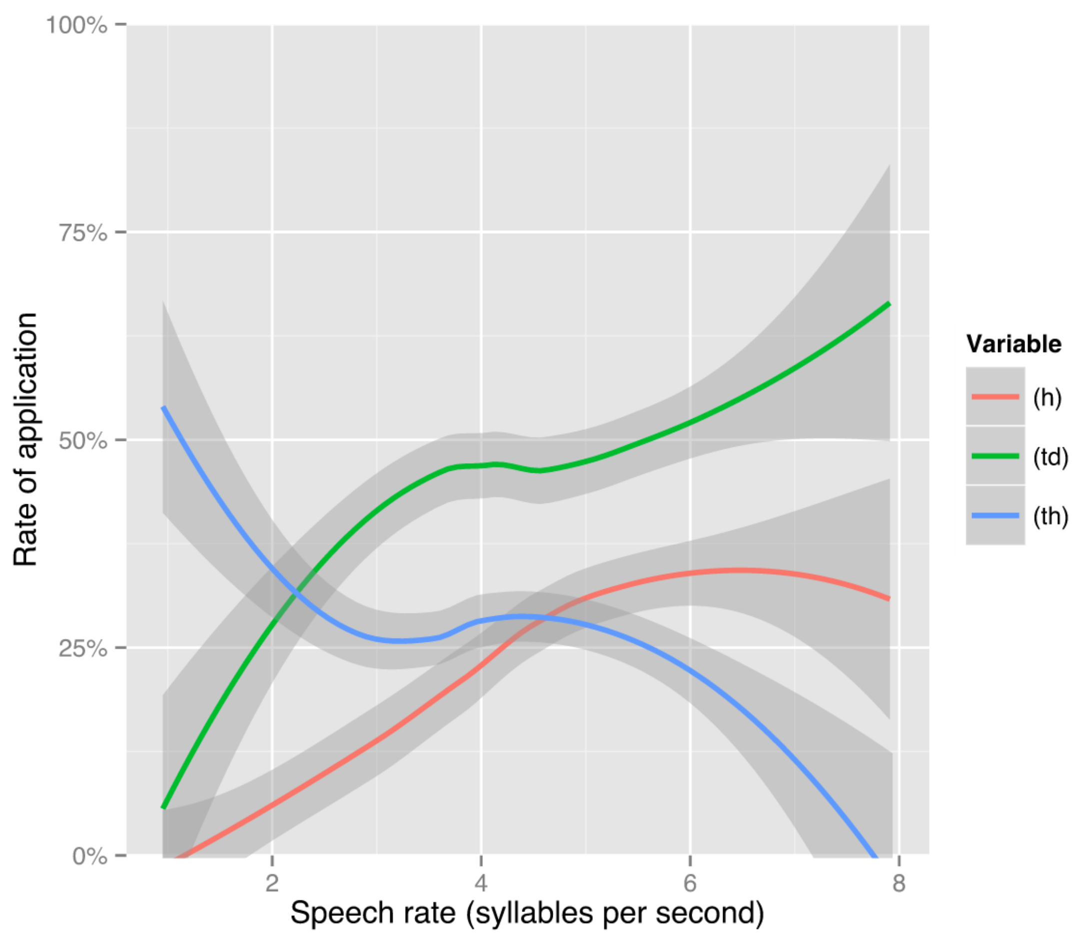
- How does this impact FAVE's accuracy in automatically identifying sociolinguistic variation?

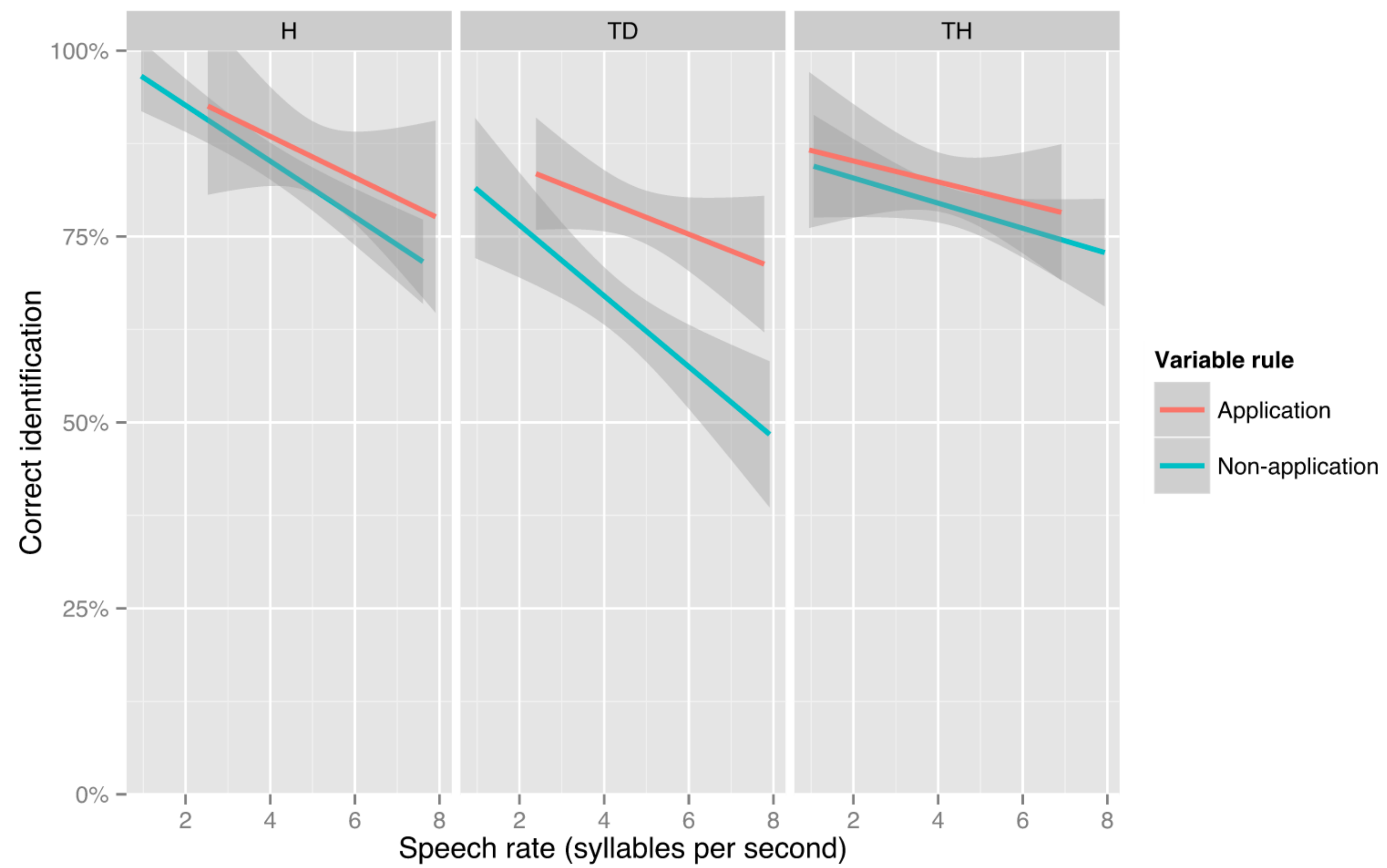




Rate of speech

- How does this impact application rates of these variable rules?





Logistic Regression

- Logistic regression models fitted for each variable using `glm`

(h)

	Estimate	Std. Error	z value	*p*	
(Intercept)	3.4867	0.6406	5.443	5.25E-08	***
application	0.4435	0.4574	0.970	0.3322	
sylls.per.s	-0.3196	0.1187	-2.692	0.0071	**

(td)

	Estimate	Std. Error	z value	*p*	
(Intercept)	1.1194	0.5087	2.201	0.0278	*
application	1.0848	0.3024	3.587	0.0003	***
voice	1.6753	0.4543	3.688	0.0002	***
sylls.per.s	-0.2457	0.1234	-1.992	0.0464	*

(th)

	Estimate	Std. Error	z value	*p*	
(Intercept)	0.41028	0.60547	0.678	0.49801	
application	1.25502	0.48978	2.562	0.0104	*
voice	1.37433	0.41584	3.305	0.001	***
sylls.per.s	-0.09837	0.10236	-0.961	0.33658	

Logistic Regression

- Logistic regression models fitted for each variable using `glm`

(h)

	Estimate	Std. Error	z value	*p*	
(Intercept)	3.4867	0.6406	5.443	5.25E-08	***
application	0.4435	0.4574	0.970	0.3322	
sylls.per.s	-0.3196	0.1187	-2.692	0.0071	**

(td)

	Estimate	Std. Error	z value	*p*	
(Intercept)	1.1194	0.5087	2.201	0.0278	*
application	1.0848	0.3024	3.587	0.0003	***
voice	1.6753	0.4543	3.688	0.0002	***
sylls.per.s	-0.2457	0.1234	-1.992	0.0464	*

(th)

	Estimate	Std. Error	z value	*p*	
(Intercept)	0.41028	0.60547	0.678	0.49801	
application	1.25502	0.48978	2.562	0.0104	*
voice	1.37433	0.41584	3.305	0.001	***
sylls.per.s	-0.09837	0.10236	-0.961	0.33658	

Logistic Regression

- Logistic regression models fitted for each variable using `glm`

(h)

	Estimate	Std. Error	z value	*p*	
(Intercept)	3.4867	0.6406	5.443	5.25E-08	***
application	0.4435	0.4574	0.970	0.3322	
sylls.per.s	-0.3196	0.1187	-2.692	0.0071	**

(td)

	Estimate	Std. Error	z value	*p*	
(Intercept)	1.1194	0.5087	2.201	0.0278	*
application	1.0848	0.3024	3.587	0.0003	***
voice	1.6753	0.4543	3.688	0.0002	***
sylls.per.s	-0.2457	0.1234	-1.992	0.0464	*

(th)

	Estimate	Std. Error	z value	*p*	
(Intercept)	0.41028	0.60547	0.678	0.49801	
application	1.25502	0.48978	2.562	0.0104	*
voice	1.37433	0.41584	3.305	0.001	***
sylls.per.s	-0.09837	0.10236	-0.961	0.33658	

Logistic Regression

- Logistic regression models fitted for each variable using `glm`

(h)

	Estimate	Std. Error	z value	*p*	
(Intercept)	3.4867	0.6406	5.443	5.25E-08	***
application	0.4435	0.4574	0.970	0.3322	
sylls.per.s	-0.3196	0.1187	-2.692	0.0071	**

(td)

	Estimate	Std. Error	z value	*p*	
(Intercept)	1.1194	0.5087	2.201	0.0278	*
application	1.0848	0.3024	3.587	0.0003	***
voice	1.6753	0.4543	3.688	0.0002	***
sylls.per.s	-0.2457	0.1234	-1.992	0.0464	*

(th)

	Estimate	Std. Error	z value	*p*	
(Intercept)	0.41028	0.60547	0.678	0.49801	
application	1.25502	0.48978	2.562	0.0104	*
voice	1.37433	0.41584	3.305	0.001	***
sylls.per.s	-0.09837	0.10236	-0.961	0.33658	

1. Introduction

Research questions

Forced alignment

Hidden Markov Models

Pronouncing dictionary

2. Methodology

Dictionary ‘hacking’

Measuring accuracy

3. Results

Overview

Detailed analysis

Rate of speech

4. Conclusion

Conclusion

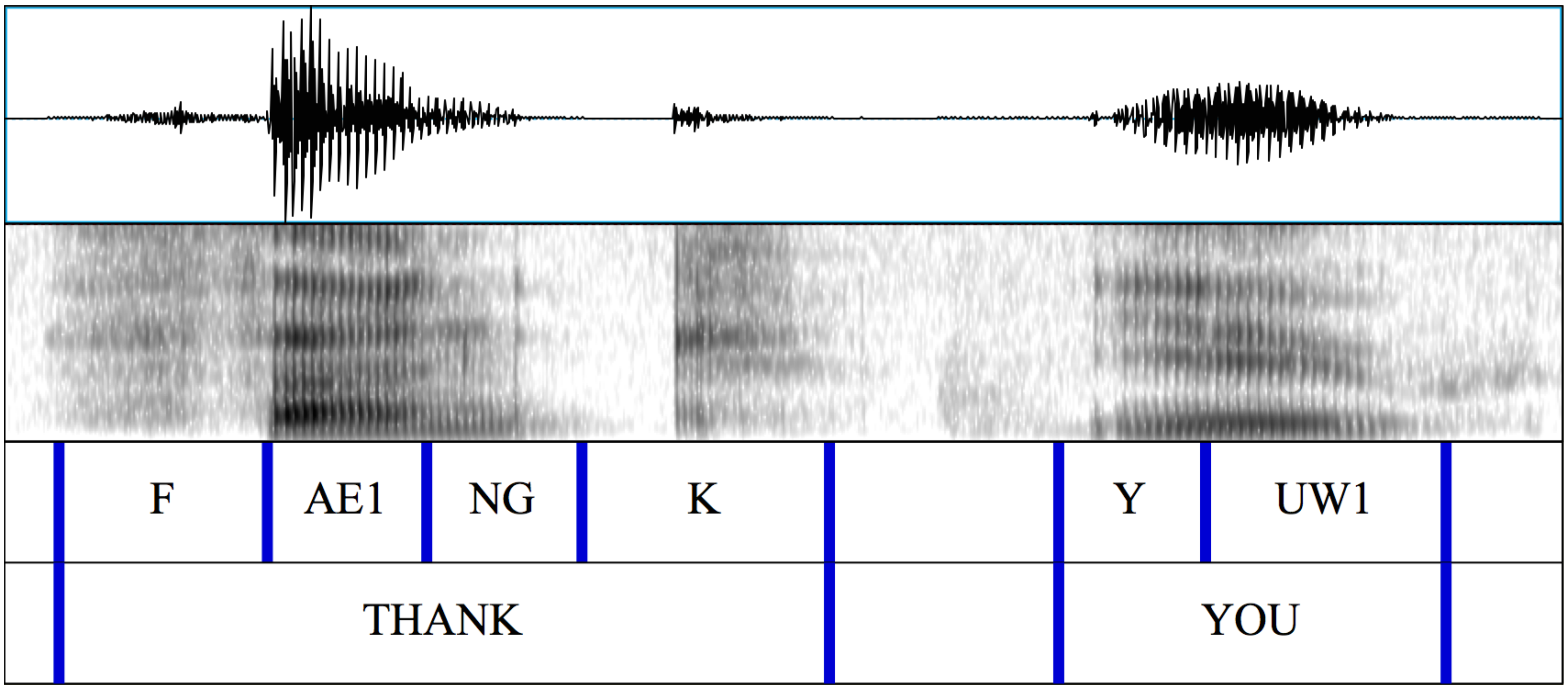
- Automated coding of phonological variation *is* possible using forced alignment
- This study has quantified the degree of error introduced by employing such a methodology
- For the most part, FAVE seems to struggle most where humans seem to struggle most!
- Reassuringly, FAVE's overall accuracy was higher for tokens where the human transcribers were in agreement (94.24%, cf. 80.92% for more ambiguous tokens)

Accuracy Efficiency

Conclusion

Thoughts for future improvement

- These tests should be carried out for a wide range of speakers and recording qualities
- Employing composite models (e.g. Yuan & Liberman 2011)
- Training speaker-specific acoustic models, or at least dialect-specific models
- Integrate some pseudo-phonology into the aligner to deal with multiple variables at once and remove the need for manual dictionary expansion


F
AE1
NG
K
Y
UW1
THANK
YOU

References

Carletta, J. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics* 22(2), 249-254.

Ghahramani, Z. 2001. An introduction to Hidden Markov Models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence* 15(1), 9-42.

Gorman, K., J. Howell, & M. Wagner. 2011. Prosodylab-Aligner: a tool for forced alignment of laboratory speech. *Proceedings of Acoustics Week in Canada,* 4-5.

Labov, W., I. Rosenfelder, & J. Fruehwald. 2013. One hundred years of sound change in Philadelphia: linear incrementation, reversal, and reanalysis. *Language* 89(1), 30-65.

MacKenzie, L., & D. Turton. 2013. Crossing the pond: extending automatic alignment techniques to British English dialect data. Presented at *New Ways of Analyzing Variation (NWAV42),* 20 October 2013.

Milne, P. 2014. *The variable pronunciations of word-final consonant clusters in a force aligned corpus of spoken French*. University of Ottawa dissertation.

Reddy, S. & J. Stanford. 2015. Toward completely automated vowel extraction: introducing DARLA. *Linguistics Vanguard.*

Rosenfelder, I., J. Fruehwald, K. Evanini, S. Seyfarth, K. Gorman, H. Prichard, & J. Yuan. 2014. *FAVE (Forced Alignment and Vowel Extraction) Program Suite,* v1.2.2 10.5281/zenodo.22281

Yuan, J. & M. Liberman. 2011. Automatic detection of “g-dropping” in American English using forced alignment. In *Proceedings of 2011 IEEE Automatic Speech Recognition and Understanding Workshop,* 490-493.

Appendix: (ing)

- Also tested 3,744 tokens of the [ɪn]~[ɪŋ]~[ɪŋg] alternation in Northern English varieties (Manchester and Blackburn) across 16 speakers
- 92.34% accurate in coding [ɪn]
- 76.74% accurate in coding [ɪŋ]
- 77.01% accurate in coding [ɪŋg]
- No human agreement rates (yet!)
    - But Yuan & Liberman (2011) report 84.9% mean accuracy rate and 86.3% human agreement rate in their comparable study of automated (ing)-coding ([ɪn] ~ [ɪŋ])