Uneven success: Racial Bias in Automatic Speech Recognition

Alicia Beckford Wassink
Department of Linguistics, University of Washington
https://depts.washington.edu/sociolab/

University of Michigan
18 January 2021
Rev. Dr. Martin Luther King, jr. Colloquium

Outline

Acknowledgements

Aims of the Talk

Background

- What do I mean by racial bias?
- Where do we see bias in language-related systems?

Methods

- Our tool: CLOx
- The sample: 4 ethnic groups from Pacific Northwest English (PNWE) study corpus
- Targeted linguistic variables

By-ethnicity results

Some surprising findings

Conclusions

The Pacific Northwest English Study

PNWE

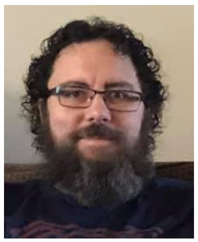
UNIVERSITY *of* WASHINGTON

DEPARTMENT OF LINGUISTICS

acknowledgements

CLOx Team:

Campion Fellin



David Nichols

Robert Squizzero

Not pictured:

Jake McManus

Amina Venton

Diana Davidson

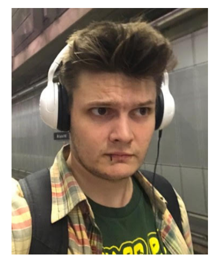
PNWE Team:

Isabel Bartholomew

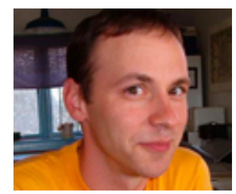
Sophia Chan

Cady Gansen

Monica Jensen



Nathan Johnson



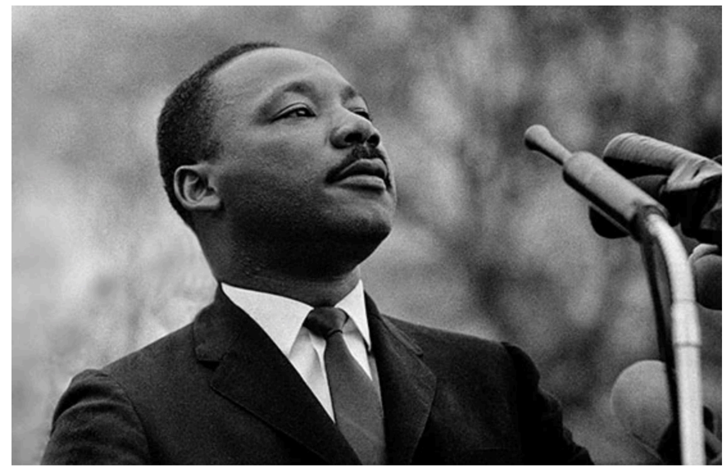
Michael Scanlon

National Science Foundation
BCS-1844350

Equality Now: the president has the power

"The new administration has the opportunity to be the first in 100 years of American history to adopt a radically new approach to the question of civil rights. It must begin, however, with the firm conviction that the principle is no longer in doubt. The day is past for tolerating vicious and inhuman opposition on a subject which determines the lives of twenty million Americans....We must decide that in a new era, there must be a new thinking. If we fail to make this positive decision, ***an awakening world will conclude that we have become a fossil nation, morally and politically; and no floods of refrigerators, automobiles or color television sets will rejuvenate our image."***

The Nation 192 (4 Feb 1961): 91-95.



Rev. Dr. Martin L. King, jr.

Rev. Ernest L. Wilson

Aims of this project

- Support for the larger PNWE research study
- Not all features of speech are handled well
- Contemporary use cases:
  - Siri, Alexa, Cortana
  - Payment-by-phone, OnStar
- Inequity in access to services
- Knowledge regarding sociolinguistic variation has yet to be exploited in acoustic model architectures
- Personal and professional significance for me: an area in which to pursue equity

Research Questions:

1. Is there a difference in error rates for four ethnicity-related subsamples?
   If so, what differences do we observe in error rate?
   What is the by-ethnicity distribution of phonetic error types?
2. What dialect features appear to be most challenging for our CLOx speech-to-text service (Microsoft)?
   Are these dialect features more typically found in the more casual speech tasks?

Background
What do I mean by racial bias?

- A form of implicit bias
  - Automatic associations or stereotypes made by individuals in the unconscious state of mind.
  - No explicit intent to harm
  - Associations influence behavior, "making people respond in biased ways even when they are not explicitly prejudiced."

National Initiative for Building Community Trust and Justice (2015)

- Defined for organizations
  - 1) Unequal access to the beneficial work of the organization, 2) Racial disparities in the structure of the organization in roles and offices, 3) Systematic pattern of inclusion and exclusion, or hierarchical distinction, in how the work proceeds, 4) Failure to examine disparities with intent to identify, address or reverse underlying causes

Maryfield (2018), Justice Research and Statistics Association

Charity Hudley (2017)

Racial bias in Linguistics?

- Language as part of the "master narrative" of cultural description
  - Linguistic categories were used to elaborate a set of cultural categories for humankind
  - Focus on languages as if these were monolithic

(Hutton, 1999)

- Classification of language groups centering a monolingual ideal
  - even sociolinguists!
  - NORMs: non-mobile, older, rural, (majority ethnicity) males

- Beliefs about who is and is not a "typical" member of a language group or speech community based upon analysts' assessment of speaker race

Colonial bias in Linguistics?

- Examining Native American language varieties only through an "endangerment lens"
  - What constitutes a native speaker?
  - What constitutes "knowing" a language?
  - Decolonized approaches to addressing language shift and language *return*

(Leonard, 2019)

- Exclusion of other varieties spoken in Native American communities (American English sociolects)

- For the PNWE study, inclusion of Yakama English allows:
  - Departure from dictum to hold certain speakers aside until after that primary work is done
  - Sophisticated study of sociolectal features (transfer from heritage language)
  - Participation in regional Pacific Northwest forms

Racial bias in Language-related technology?

Koenecke, et al. (2020)

- Contemporaneous with the PNWE ASR study, Stanford study of Word Error Rates (WERs) in sociolinguistic corpora of AAE speech
  - 5 ASR systems (Google, Amazon, Apple, IBM, Microsoft)
  - only previous sociolinguistic study of racial bias in ASR system performance
  - Syntactic constructions (copula deletion "He a pastor.")
- Examination of *perplexity:*
  - <u>Def.:</u> In language models, the number of reasonable continuations of a phrase
  - Language model <u>not</u> prone to bias (perplexity *lower* for AAE than GAE), even though high WERs were observed.
- *Results "must be due" to phonetic factors*

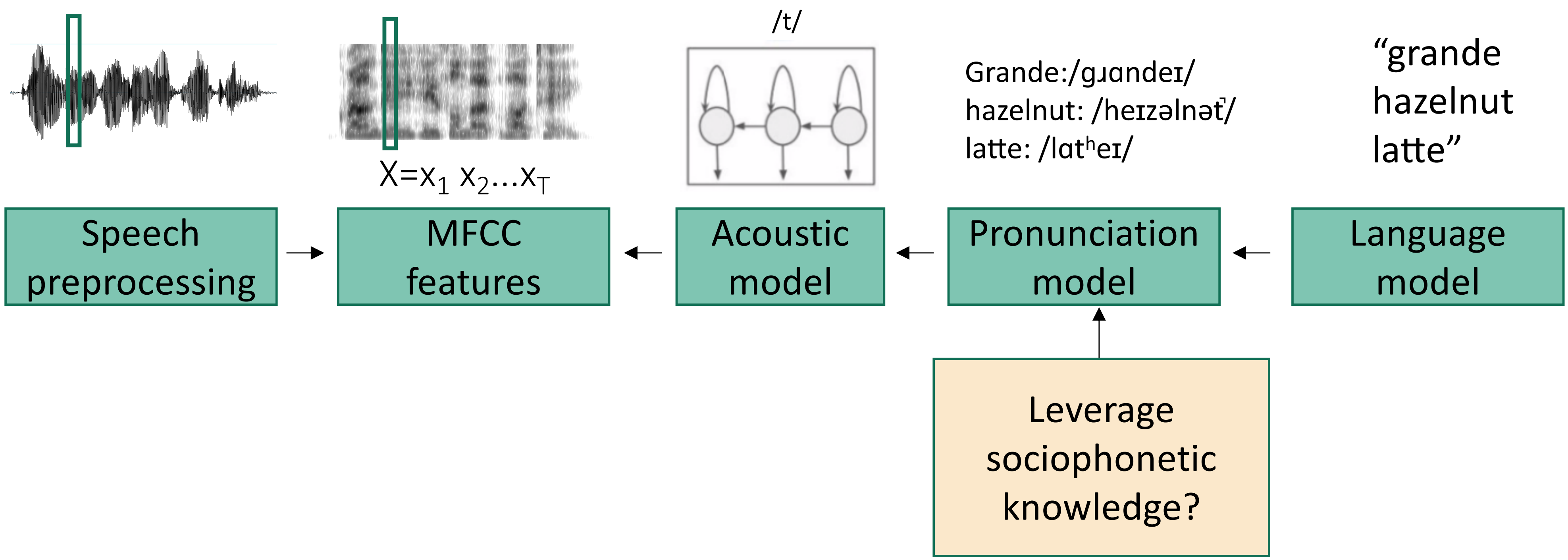
Ex. "the dog jumped over the______."

Fence
Box
Stick

Perplexity=3

Speech Recognition: primer

- Black box problem, but architecture is probably something like …



Adapted from Hui (2019)

Methods

Talkers

16 speakers, 4 Ethnic groups

Yakima (4 M, 2 F)
Mexican American (2 M, 1 F)
African American (1 M, 2 F)
Caucasian American (1 M, 3 F)

Note: Speaker classification into ethnic groups was based upon:

- Speaker's self-identification
- Social network data (membership in a speech community)
- Length of time in speech community

Data amounts

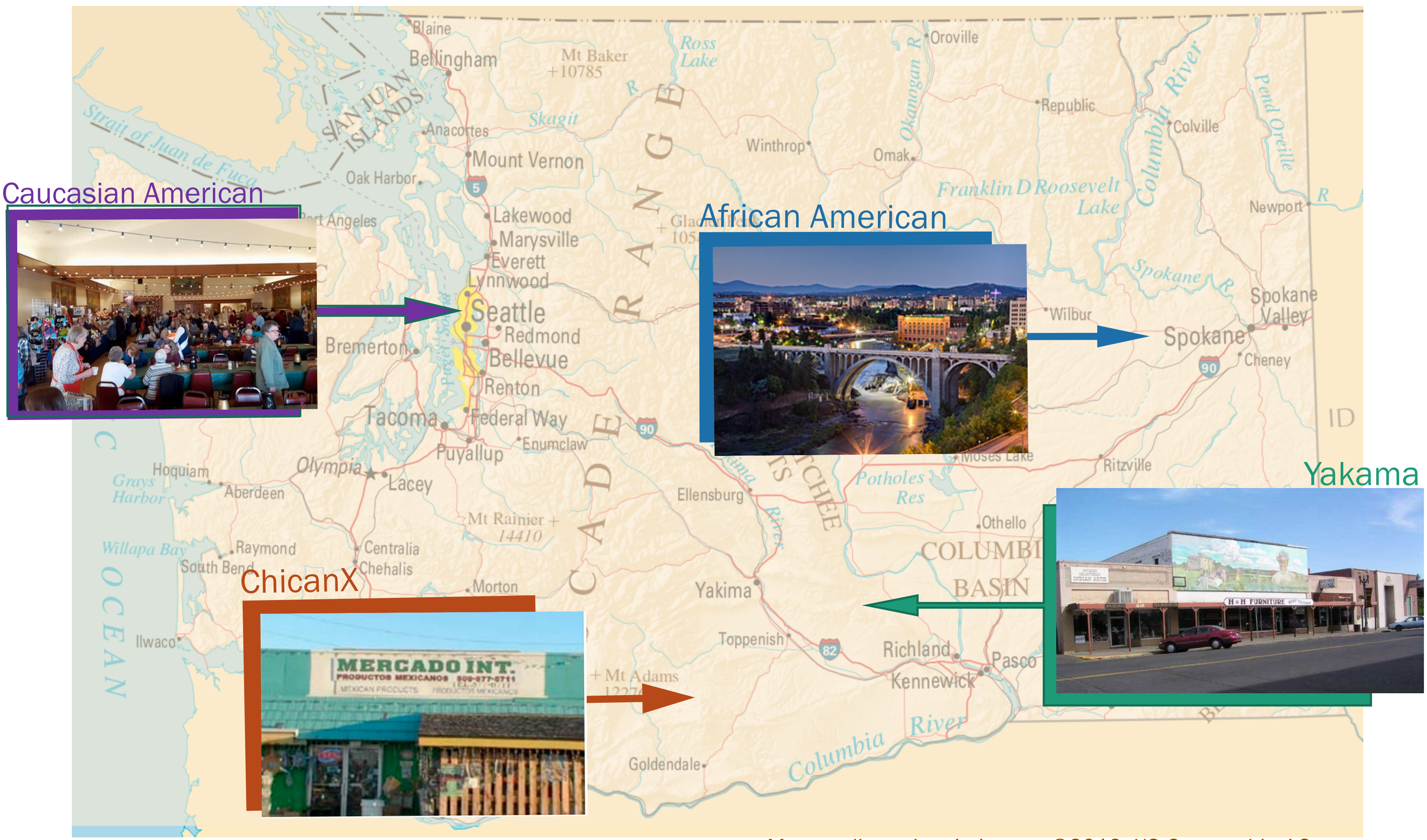
Approx. 45 - 90 min. of speech per recording
Minimum of 20 min. of speech per talker
9,174 - 22,773 words per ethnic group

Corpus

13 hours (4.99 GB)

The Pacific Northwest English Study
PNWE
UNIVERSITY of WASHINGTON
DEPARTMENT OF LINGUISTICS

Speaker sample: 4 WA dialects



Map credit: nationalatlas.gov ©2019: US Geographical Survey

Tasks

Three tasks:

Task	Style	Common Lexical content?	Task Word Count
Free-flowing speech	Casual (dyadic)	Uncontrolled (common topics, QGenII)	517-6019
Lexical Task*	Semi-casual (individual)	Semi-Controlled	218-691
Reading passage “The Cat and the Mice” (Aesop’s Fables)	Citation	Controlled	342 (fixed)

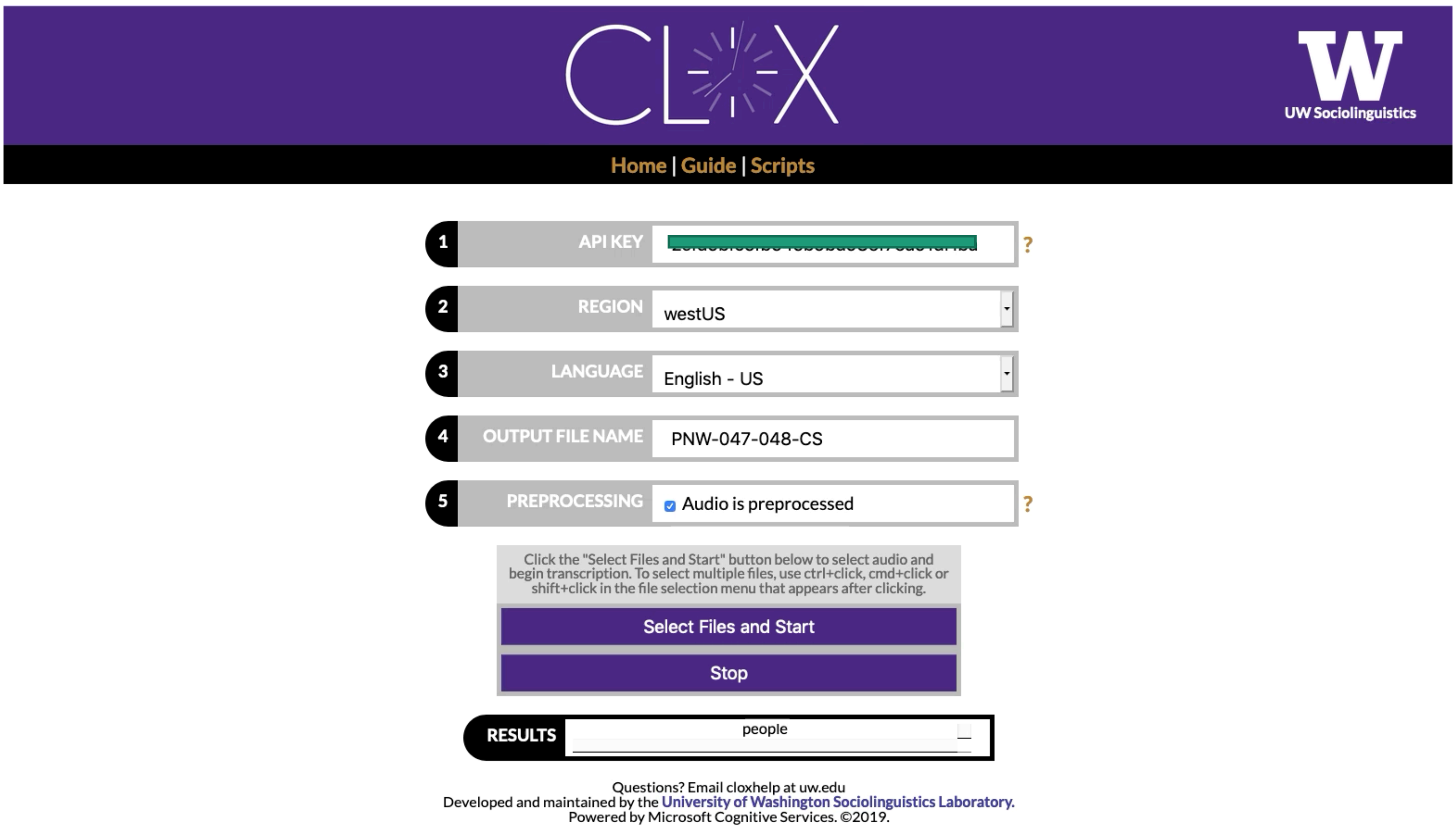
17common variables

<u>Lexical task (word games):</u>
Lists (numbers, days of the week, breakfast foods, farm animals)
Minimal pairs (dawn/don)
Semantic differentials (what is the difference in meaning between a “sack” and a “bag”?)

Our Tool: CLOx

- Client Libraries Oxford
- Automated audio transcription service for linguists developed by the Sociolinguistics Laboratory at the University of Washington.
- Automatic speech recognition uses the Speech-to-text service SDK (Microsoft Cognitive Services, Speech Division).
- CLOx delivers a conversational recording to MS, which returns plain-text transcribed output, then CLOx performs output checking and supplies timestamps indicating the start and end time of each run of speech.

Our Tool: CLOx



Data Handling

- All recordings submitted to ASR tool (CLOx)
- Transcripts returned by CLOx were manually coded for errors
  - Each recording was audited using ELAN, errors manually entered into an Excel database
  - Erroneous phone
  - Intended phone
  - Inter-rater reliability (agreement in coding over 20% of each file)

1	Text	Onset	Offset	Erroneous To	Corrected Token	Token Class	Analyst	Comment	TokenClass	Count
2	What's the opposite of friends back The opposite of positive negative	1.97	6.2						ing	0
3	And with the kind of dessert that's often served at birthdays or weddings NULL	6.84	10.43	NULL	cake	NULL			TH	0
4	What's the difference between that and pie	11.96	13.4						?	0
5	My husband prefers pie Pie tends to be a top and a bottom crust or sometimes without a top crust with fruit or something in the middle and then cake is just flour and sugar concoction baked all the way through	15.07	29.14						ɹ	0
6	All the way through Excuse me usually with frosting	29.75	32.6						d	2
7	So far everyone I've interviewed has purred pie I prefer cake myself mean to my husband says no make me a birthday pie	33.94	42.71						CC	0
8	Actually went to a winning ones where it was like potluck pie Oh that's fun Yeah that would be fun	44.4	49.32						l	0
9	I know who really definitively has the best Apple pie recipe	50.17	53.33						ɪ	0
10	OK when someone speaking too generally not giving enough details there being too vague If you're hungry between meals you might fix yourself a snack or some kinds of foods that people have for sex	56.77	68.24						ɔ	0
11	If they're healthy sort of people though rabbit piece of fruit or some grapes Things like that Most of the rest of us go for chips and	69.42	77.03	rabbit	grab a	O		initial cluster simplification; V in rhyme; C in coda	æg	0
12	Pretzels and what NULL I usually have	78.91	81.23	NULL	would	NULL			æ	0
13	Hum	82.08	82.64						ɛg	2
14	My daughter goes through the bread drawer and start just eats pieces of bread	83.93	87.11						ʌ	0
15	Drive	87.89	88.34	Drive	dry	O			ow	0
16	What kind of fruits grapes apples bananas	89.96	93.7						prel	0

Phonetic Error Rate (PER)

Normalized frequency measure, calculated as the proportion of all errors falling into a particular sociolinguistic variable class

E	**Erroneous forms across all targeted linguistic variables in a corpus**
N	**Total word count for the corpus**
B	**Base of normalization = 100 words**
nf	**(E/N)*B** **Number of error in corpus / total corpus x base of normalization**

E = 668
N = 16,276
nf = (668/16276) *100
= 4.104

General error types

Code	Label	Example error	Target	IPA
R	reduction	lotta	lot of	varies
D	disfluencies	enough	and uh	
NC	no code	changing	digging	
NULL	words inserted	could ("windows <u>could</u> they would")	ø	
PN	Proper name	topless	Toppenish	
H	Homophone	are~R~our	are~R~our	

- Not associated with any specific dialect
- Not targeted for sociophonetic study

Targeted Sociolinguistic Variables

Consonants:

Wassink (2017), Wassink and Hargus (2020)

Code	Sociolinguistic Label	Example error	Target	IPA
(ing)	-ing (unstressed)	pick into	picking too	[ɪŋ] vs [ɪn] vs [in]
(TH)	th-stopping	den	then	/ð/ → [d]
(ʔ)	word-medial glottalization	right are	writer	/t/ → [ʔ]
(ɹ)	coda-r deletion	what a	water	/ɹ/ → Ø
(d)	consonant cluster deletion	pace [peɪs]	paced /peɪst/	/st/ → [s]
(l)	lenition	sheep	cheap	/tʃ/ → [ʃ]

<u>Why a common set of variables?</u>

- Assess extent to which regional changes present a problem for ASR
- We know that some forms span non-standard dialects of English
- It may be that certain errors are particular to certain sociolects
- If we see common errors for multiple groups, inclusion in the AM will represent greater gains for ASR.

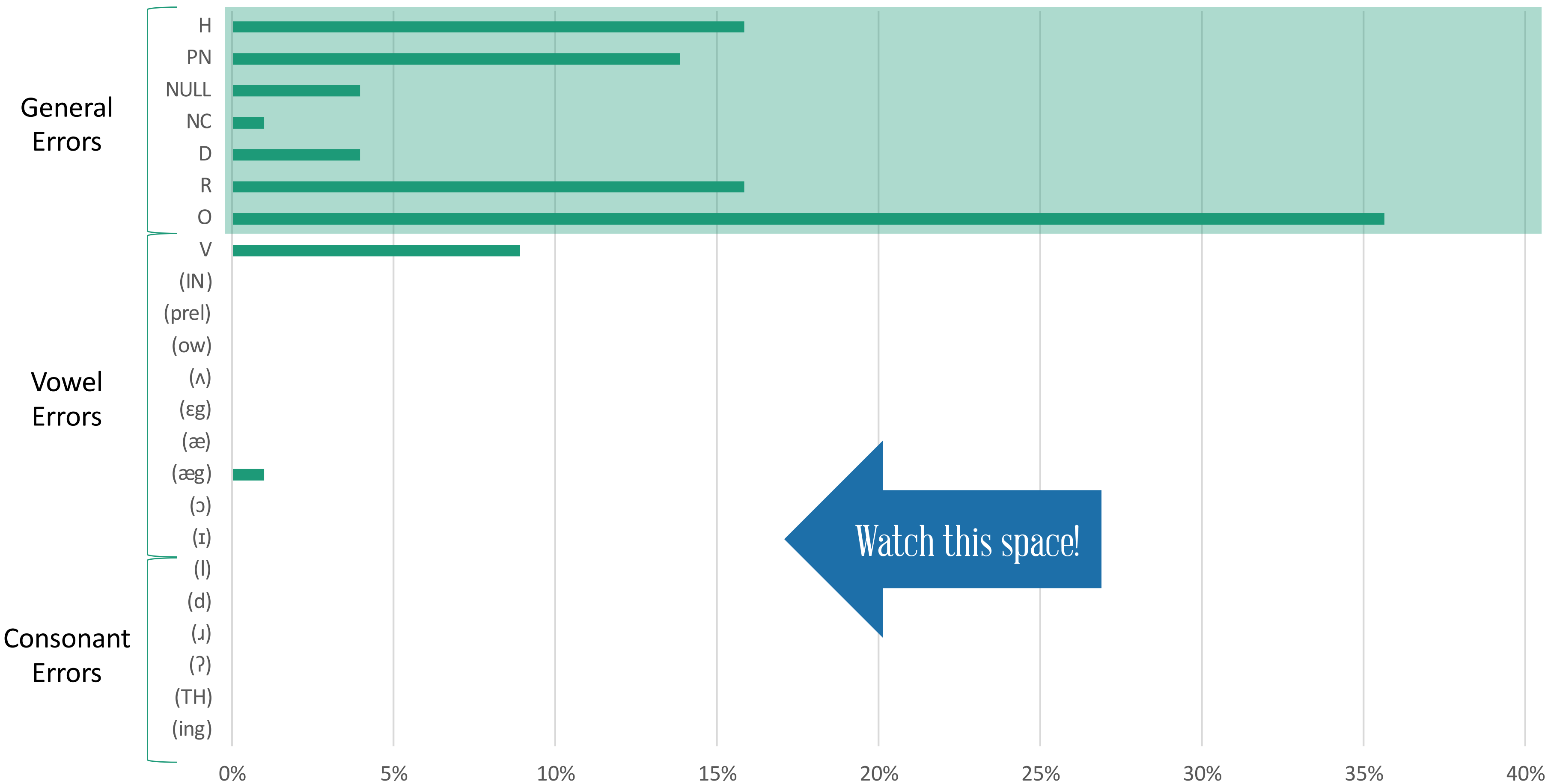
Sociolinguistic Variables

Vowels:

Code	Sociolinguistic Label	Example error	Target	IPA
(ɪ)	(ɪ)-tensing	peaking	picking	/ɪ/⟶ [i]
(ɔ)	caught/cot merger	com, cot	calm, caught	/ɔ/ ⟶ [a], /ɔ/ ⟶ [ɑ]
(æg)	pre-voiced velar (æ)-raising	beg	bag	/æg/ ⟶ [e:g]
(æ)	mistaking (æ) for other Vowel	infect	in fact	/æ/⟶ [a], /æ/⟶ [ɛ]
(ɛg)	pre-voiced velar (ɛ)-raising	beg	bake	/ɛg/ ⟶ [e:g]
(ʌ)	(ʌ)-raising	is	us	/ʌ/ ⟶ [ɨ], /ʌ/ ⟶ [ɪ]
(ow)	(ow)-fronting	boot	boat	/ow/⟶ [u]
(prel)	prelateral back vowel merger	full, hole	fool, hull	/ul/⟷ /ol/, /ʊl/⟷ /ul/, /ʌl/⟷ /ol/
(IN)	pin/pen merger	pin	pen	/ɪn/⟷ /ɛn/
V	other vowel error	greet	great	varies
O	other (phonetic/phonological errors)	thing, faults	vague, false	varies

- <u>ARE</u> associated with specific dialects
- <u>ARE</u> targeted for sociophonetic study

CLOx Errors, by type (Caucasian American Subsample)



	(ing)	(TH)	(ʔ)	(ɹ)	(d)	(l)	(ɪ)	(ɔ)	(æg)	(æ)	(ɛg)	(ʌ)	(ow)	(prel)	(IN)	V	O	R	D	NC	NULL	PN	H
%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	9%	36%	16%	4%	1%	4%	14%	16%

Results

RQ1: Is there a difference in error rates between four ethnicity-related subsamples?

Yes!

- Overall *nf*, by ethnicity

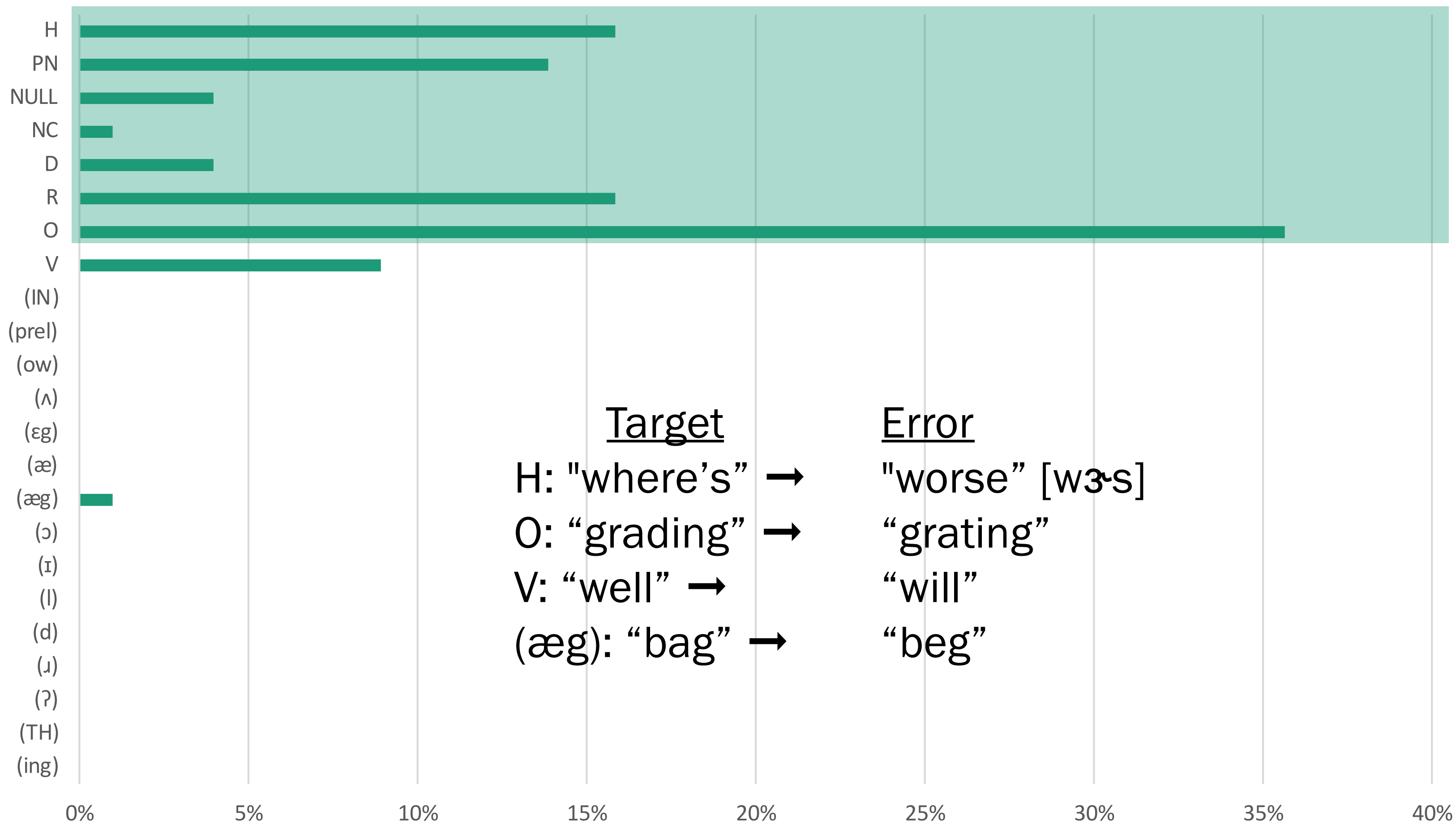
Group	N=	*nf*
Caucasian American	19,142	1.6
African American	22,773	3.6
Yakama	22,695	6.3
ChicanX	9174	6.6

One-Way ANOVA ($F(3, 788)=4.514$, $p<0.001$). Tukey's HSD: Yakama~Caucasian-Am (p=0.04) Caucasian-Am~ChicanX (p=0.00)

#1: Fewest errors (*nf=1.6*)

What is the by-ethnicity distribution of phonetic error types?

CLOx Errors, by type (Caucasian American Subsample)

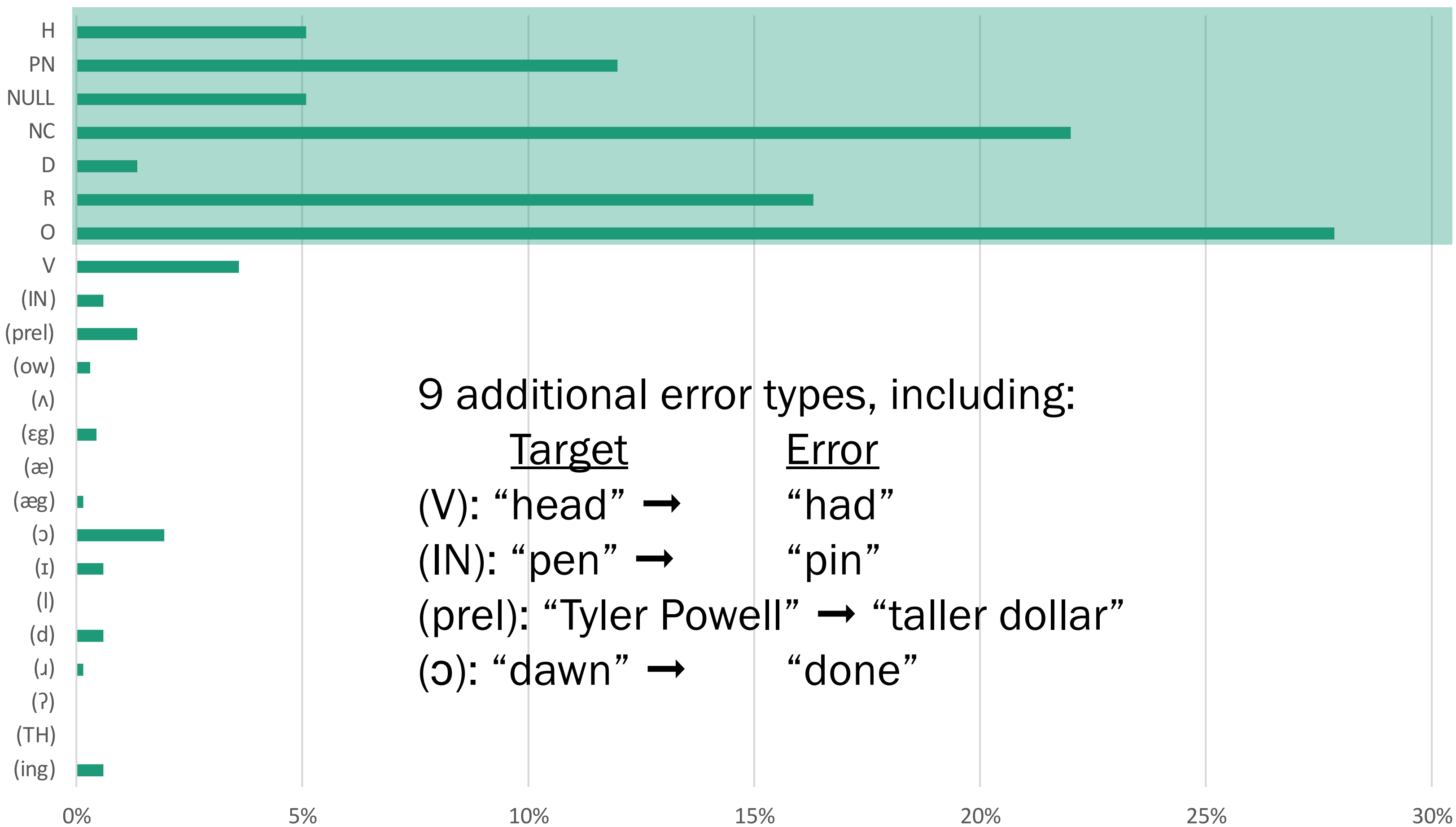


	Target		Error
H:	"where's"	➞	"worse" [wɝs]
O:	"grading"	➞	"grating"
V:	"well"	➞	"will"
(æg):	"bag"	➞	"beg"

	(ing)	(TH)	(ʔ)	(ɹ)	(d)	(l)	(ɪ)	(ɔ)	(æg)	(æ)	(ɛg)	(ʌ)	(ow)	(prel)	(IN)	V	O	R	D	NC	NULL	PN	H
%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	9%	36%	16%	4%	1%	4%	14%	16%

#2: (*nf=3.6*)

CLOx Errors, by type (African American Subsample)



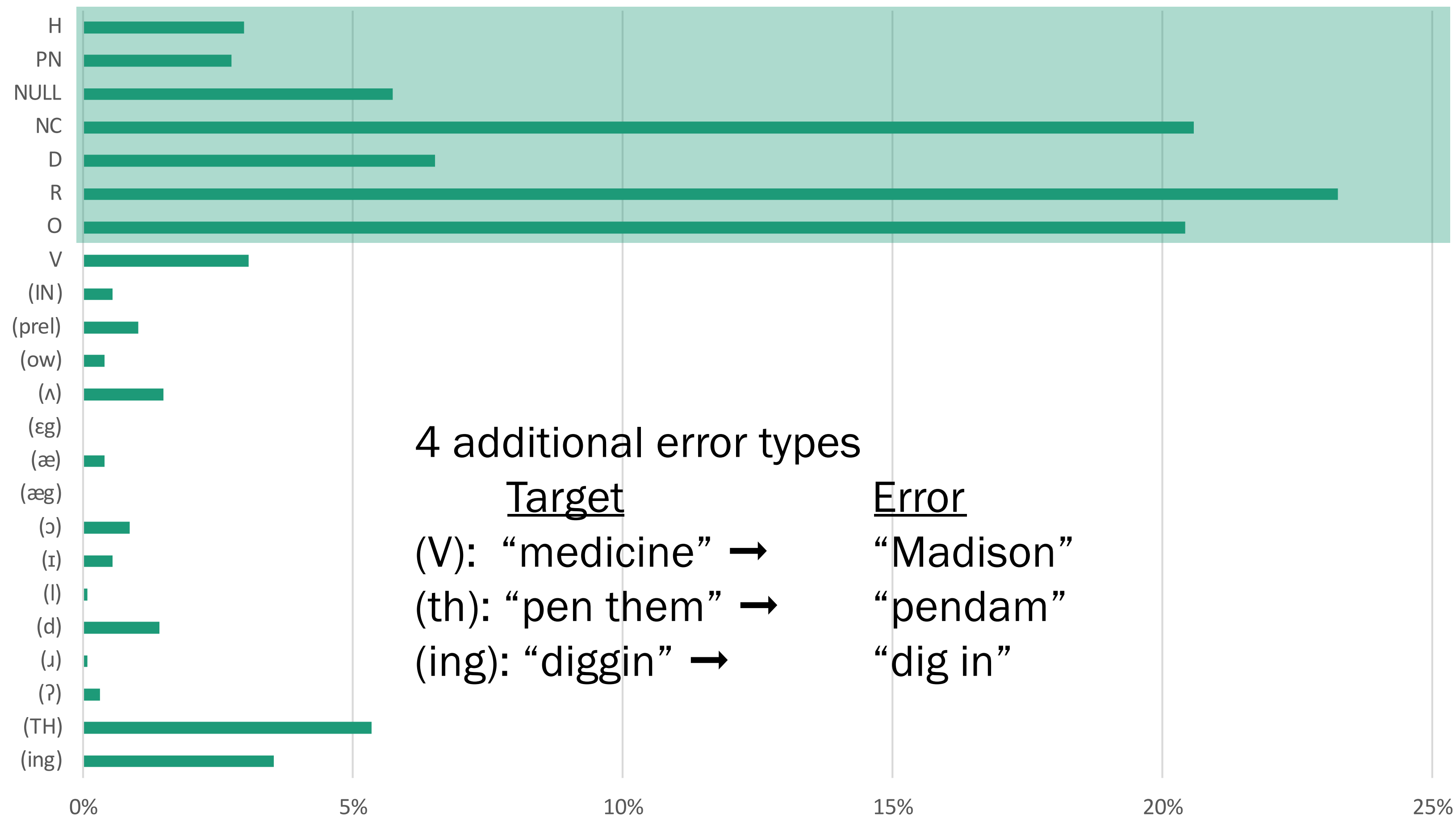
9 additional error types, including:

	Target	Error
(V):	"head" ➞	"had"
(IN):	"pen" ➞	"pin"
(prel):	"Tyler Powell" ➞	"taller dollar"
(ɔ):	"dawn" ➞	"done"

	(ing)	(TH)	(ʔ)	(ɹ)	(d)	(l)	(ɪ)	(ɔ)	(æg)	(æ)	(ɛg)	(ʌ)	(ow)	(prel)	(IN)	V	O	R	D	NC	NULL	PN	H
%	1%	0%	0%	0%	1%	0%	1%	2%	0%	0%	0%	0%	0%	1%	1%	4%	28%	16%	1%	22%	5%	12%	5%

#4: *(nf=6.3)*

CLOx Errors, by type (Yakama Subsample)

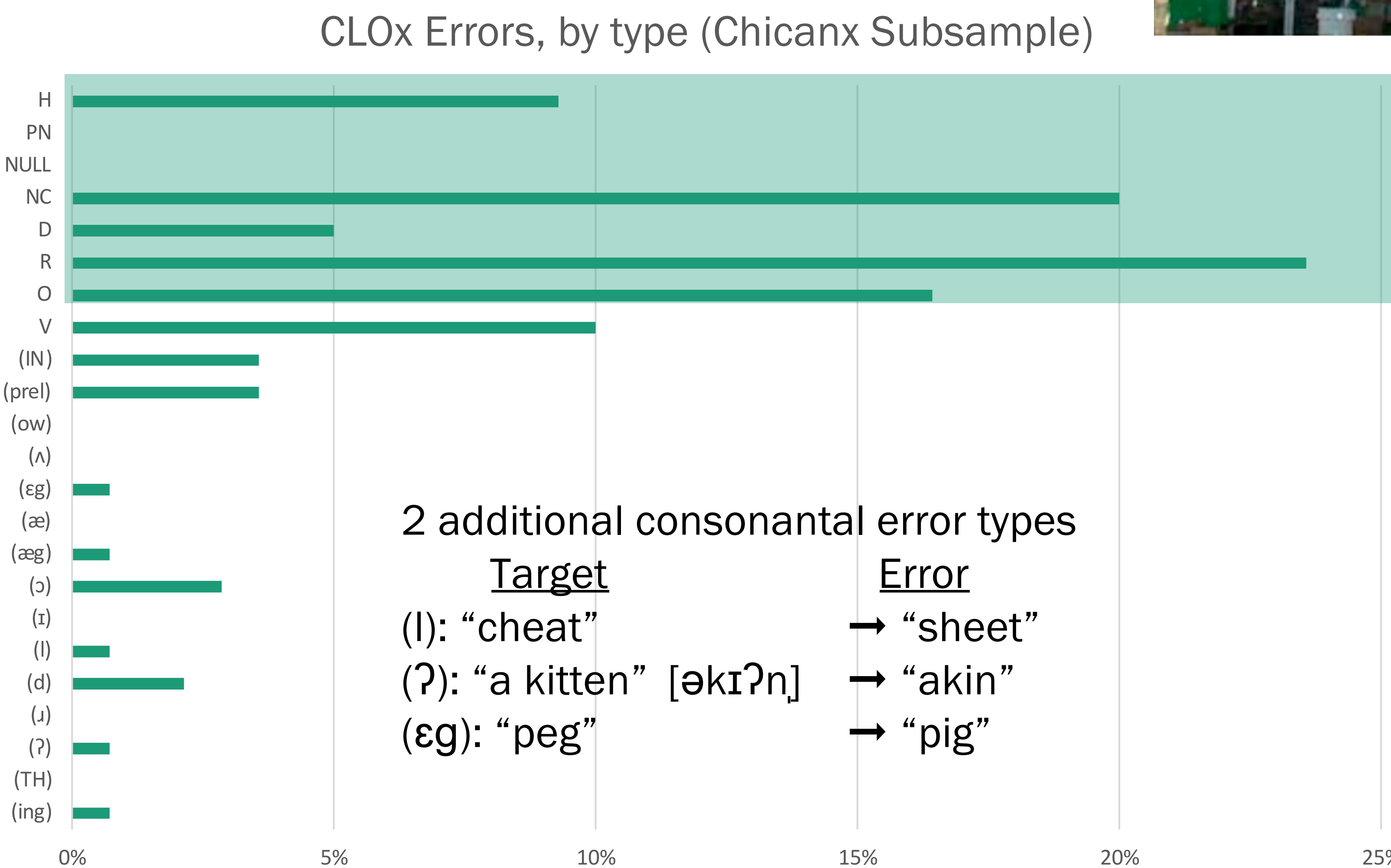


	(ing)	(TH)	(ʔ)	(ɹ)	(d)	(l)	(ɪ)	(ɔ)	(æg)	(æ)	(ɛg)	(ʌ)	(ow)	(prel)	(IN)	V	O	R	D	NC	NULL	PN	H
%	4%	5%	0%	0%	1%	0%	1%	1%	0%	0%	0%	1%	0%	1%	1%	3%	20%	23%	7%	21%	6%	3%	3%

4 additional error types

	Target	Error
(V):	“medicine” ➞	“Madison”
(th):	“pen them” ➞	“pendam”
(ing):	“diggin” ➞	“dig in”

#3: (*nf=6.6*)



	(ing)	(TH)	(ʔ)	(ɹ)	(d)	(l)	(ɪ)	(ɔ)	(æg)	(æ)	(ɛg)	(ʌ)	(ow)	(prel)	(IN)	V	O	R	D	NC	NULL	PN	H
%	1%	0%	1%	0%	2%	1%	0%	3%	1%	0%	1%	0%	0%	4%	4%	10%	16%	24%	5%	20%	0%	0%	9%

2 additional consonantal error types

Target	Error
(l): “cheat”	➞ “sheet”
(ʔ): “a kitten” [əkɪʔn̩]	➞ “akin”
(ɛg): “peg”	➞ “pig”

By-Task Results

What dialect features appear to be most challenging for our CLOx speech-to-text service (Microsoft)?
Are these dialect features more typically found in the more casual speech tasks?

70% of errors *were* in CS

Mean PER for lexical task materials was relatively low

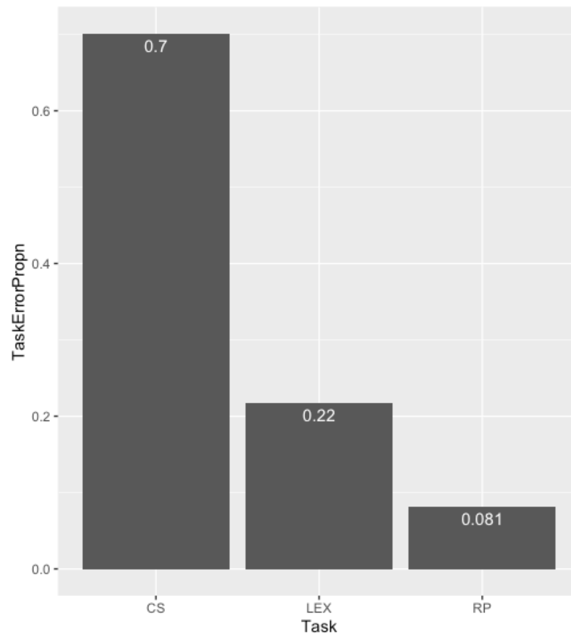


Figure 2. Errors, by Task. All groups pooled. CS=Conversational Speech, LEX=Lexical Task, RP=Reading Passage

Which sociolinguistic variables were most problematic for the MS ASR system?

(th)-stopping
/ɔ/ vs. /ɑ/
CC simplification
Prelateral merger of /ul/~/ʊl/

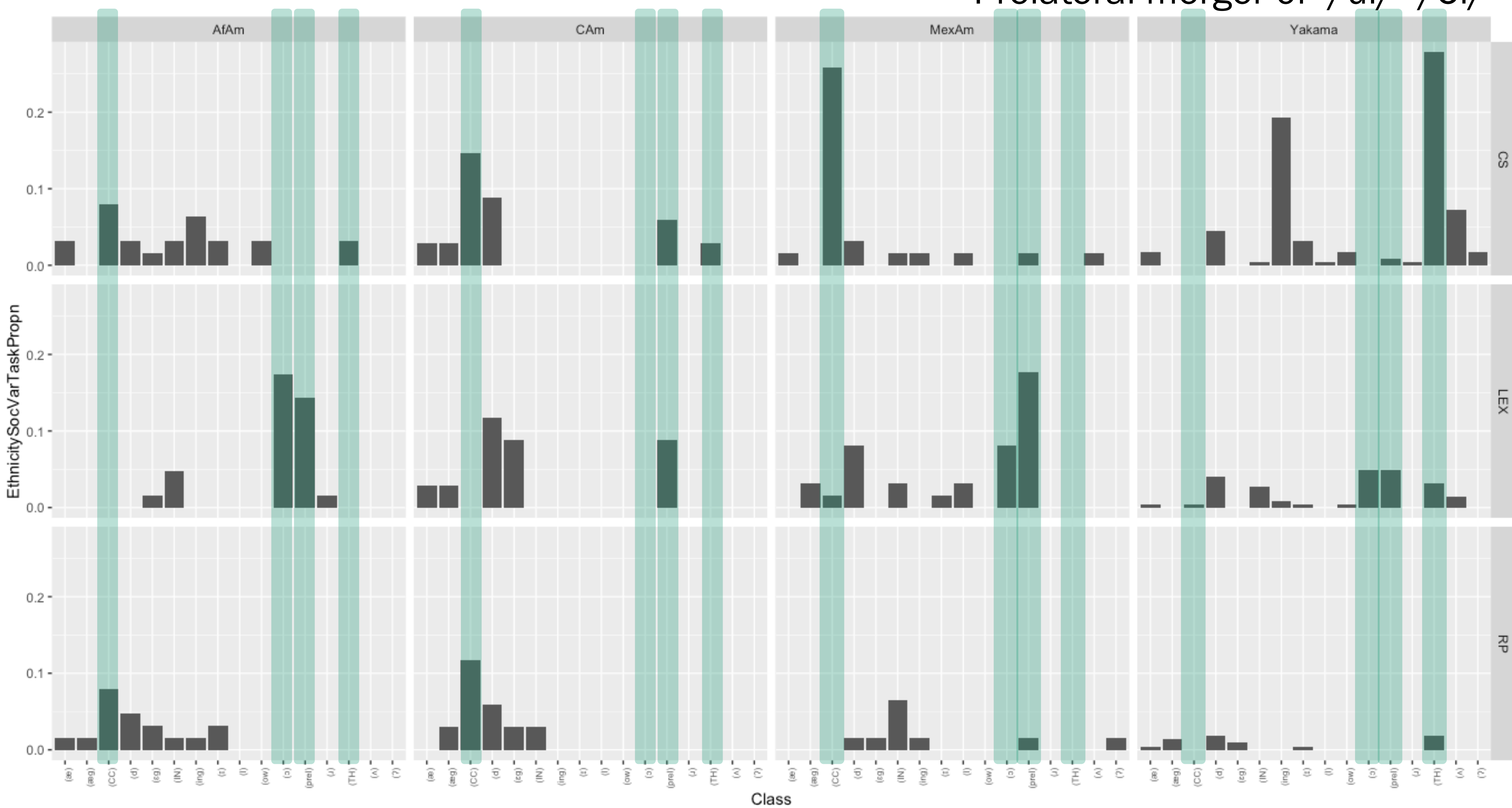


Figure 4. PER, by Sociolinguistic variable Class, Task, and Ethnicity.

(æ)	(æg)	(CC)	(d)	(ɛg)	(IN)	(ing)	(ɪ)	(l)	(ow)	(ɔ)	(prel)	(ɹ)	(TH)	(ʌ)	(ʔ)

Conclusions *and*
Where do we go from here?

This research has accomplished a cross-ethnicity comparison of dialect-based ASR performance

- Important! Quantified contribution of linguistic variables to error profile
- It's worth it! Eliminate approximately 26% of observed errors
- ASR is a useful tool on the way to "actual" linguistic analysis.

Where does the PNWE team go from here?

- Collaborate on and advocate for leveraging sociolinguistic knowledge of the fine phonetic detail in dialect variation
- Working on new pronunciation model that implements 15 of our targeted sociolinguistic variables
- Building ASR service using freely-available Kaldi architecture

Conclusions *and* Where do we go from here?

Where can linguists go from here? Some ideas:

- With respect to analysis of sociolectal variation, we need:
  - Further work on *variation* in AAE and other sociolectal varieties
  - Methods for study of multilectal speech
  - More expansive notion of native speaker
- Undoing racial and colonial bias:
  - “Look out for the overlooked”
  - Who gets excluded from linguistic research?
  - Address organizational role-related disparities (employment, tenure and promotion)

“Look out for The Overlooked”
-- folk saying, popularized recently by Kamala Harris in The Truths We Hold (2019)

Thank you!

wassink@uw.edu

Perception Test: https://depts.washington.edu/sociolab

CLOx: https://clox.ling.washington.edu/

References

Biadsy, Fadi, Soltau, Hagen, Mangu, Lidia, Navratil, Jiri, and Hirschberg, Julia (2010) Discriminative Phonotactics for Dialect Recognition Using Context-Dependent Phone Classifiers. Odyssey 2010, Jun 10-Jul 1

Charity Hudley, A. H. (2017). Language and racialization. The Oxford Handbook of Language and Society. Oxford, UK: Oxford Handbooks.

Harris, Kamala (2019) *The Truths We Hold: an American Journey.* Penguin Books.

Hui, Jonathan (2019) "Speech Recognition — GMM, HMM", The Medium.com accessed online 1/11/2021

Hutton, C. (1999). Linguistics and the Third Reich: Mother-Tongue Fascism, Race, and the Science of Language. London: Routledge.

King, Martin L. (1961) "Equality Now: the President Has the Power" *The Nation* 192 (4 Feb 1961): 91-95.

Koenecke, Allison, Nam, Andrew, Lake, Emily, Nudell, Joe, Quartey, Minnie, Mengesha, Zion, Toups, Connor, Rickford, John R., Jurafsky, Dan, & Goel, Sharad (2020) Racial disparities in automated speech recognition, *Proc. of the National Academy of Sciences, 117(14),* April 7: 7684-7689.

Leonard, W. Y. (2019) Musings on Native American Language Reclamation and Sociolinguistics. *Items.* SSRC.

Maryfield, Bailey (2018) Implicit Racial Bias. Justice Research and Statistics Association.

National Initiative for Building Community Trust and Justice. (2015). Implicit bias. Community-Oriented Trust and Justice Briefs. Washington, DC: Office of Community Oriented Policing Services.

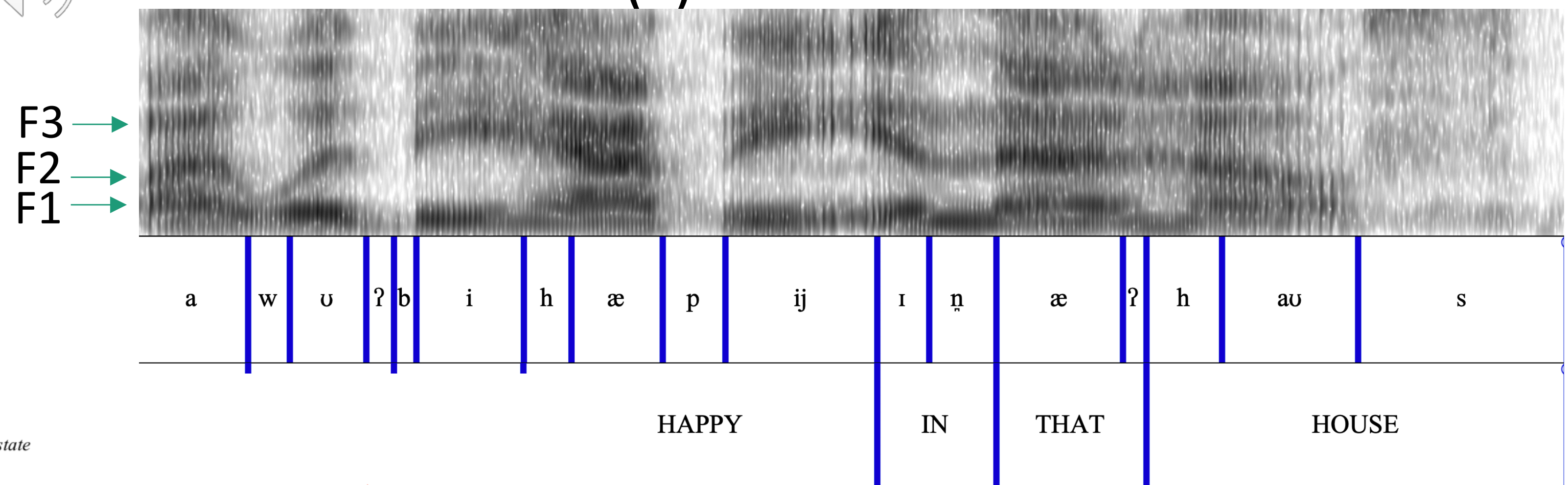
Wassink, Alicia (2017) The Vowels of Washington State In Speech in the Western States: vol. 1: the coastal states (Fridland, V. Kendall, T. Wassink, A.B. and Evans, B., eds.). Publications of the American Dialect Society. Duke University Press.

Wassink, Alicia, Gansen, Cady, and Bartholomew, Isabel (2020, unpublished ms) Uneven success: automatic speech recognition and ethnicity-related dialects, submitted to *Speech Communication.*

Wassink, A.B. and Hargus, S. (2020) "Heritage Language and Features and the Yakima English Dialect". In, Speech in the Western States, vol 3: understudied dialects (V. Fridland, A. Wassink, T. Kendall and L. Hall-Lew, eds). Publications of the American Dialect Society 103. Durham: Duke UP.

Reading Passage example

African American (F)



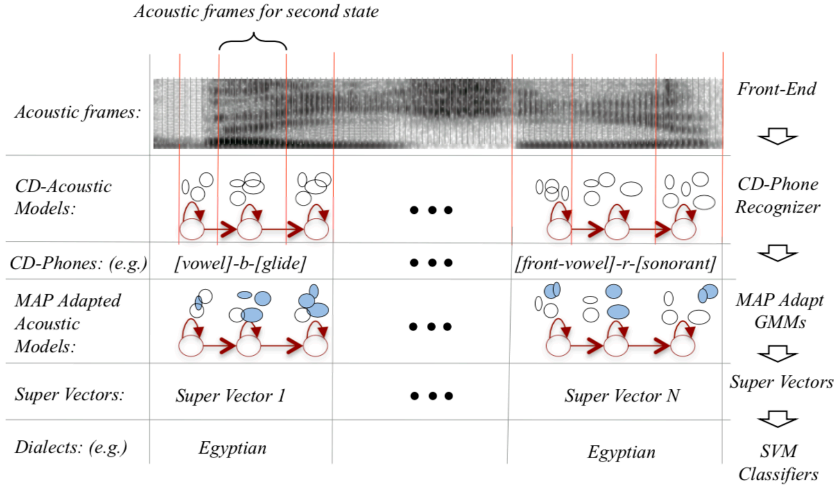


Figure 1: Dialect Classification of Context-Depdendent Phones

Source: Biadsy et al. (2010)

Within subsample ANOVA tests of mean difference in PER, by Task

	Estimate	Std. Error	t value	Pr(>|t|)
African American				
(Intercept)	0.023531	0.005538	4.249	6.59e-05 ***
TaskLEX	-0.012665	0.007832	-1.617	0.1104
TaskRP	-0.016262	0.007832	-2.076	0.0416 *
	F-statistic: 2.379 on 2 and 69 DF, p-value: 0.1002			
Caucasian American				
(Intercept)	0.027419	0.006028	4.549	2.25e-05 ***
TaskLEX	-0.017608	0.008525	-2.065	0.04264 *
TaskRP	-0.022984	0.008525	-2.696	0.00881 **
	F-statistic: 3.978 on 2 and 69 DF, p-value: 0.02318			
Yakama				
(Intercept)	0.032561	0.006630	4.911	5.84e-06 ***
TaskLEX	-0.025233	0.009376	-2.691	0.00892 **
TaskRP	-0.030782	0.009376	-3.283	0.00161 **
	F-statistic: 6.124 on 2 and 69 DF, p-value: 0.003562			
Mexican American				
(Intercept)	0.028016	0.005270	5.316	1.23e-06 ***
TaskLEX	-0.017346	0.007453	-2.328	0.02288 *
TaskRP	-0.025036	0.007453	-3.359	0.00128 **
	F-statistic: 5.922 on 2 and 69 DF, p-value: 0.004229			